SPERRY RAND

UNIVAC

# PROSODIC AIDS TO SPEECH RECOGNITION

by

Wayne A. Lea
Mark F. Medress
Toby E. Skinner

Defense Systems Division
St. Paul, Minnesota
(612-456-2430)



Semiannual Technical Report Submitted To:

Advanced Research Projects Agency
1400 Wilson Boulevard
Arlington, Virginia 22209

Attention: Dr. L. G. Roberts

Report No. PX 7940

1 October 1972

This research was supported by the Advanced Research Projects Agency of the Department of Defense under Contract No. DAHC15-72-C-0138, ARPA Order No. 2010, Program Code No. 90536. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the Advanced Research Projects Agency or the U. S. Government.

Unclassified
Security Classification

## DOCUMENT CONTROL DATA - R & D

*(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)*

1. ORIGINATING ACTIVITY (Corporate author)
Univac Defense Systems Division
P.O. Box 3525
St. Paul, Minnesota 55165

2a. REPORT SECURITY CLASSIFICATION
Unclassified

2b. GROUP

3. REPORT TITLE
Prosodic Aids to Speech Recognition

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)
Semiannual Technical Report; 1 March-31 August, 1972

5. AUTHOR(S) (First name, middle initial, last name)
1) Wayne A. Lea
2) Mark F. Medress
3) Toby E. Skinner

| 6. REPORT DATE | 7a. TOTAL NO. OF PAGES | 7b. NO. OF REFS |
|---|---|---|
| 1 October 1972 | 68 | 94 |

| 8a. CONTRACT OR GRANT NO. | 9a. ORIGINATOR'S REPORT NUMBER(S) |
|---|---|
| DAHC15-72-C-0138 | Univac Report No. PX 7940 |
| b. PROJECT NO. | |
| c. | 9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report) None |
| d. | |

10. DISTRIBUTION STATEMENT
Distribution of this document is unlimited.

| 11. SUPPLEMENTARY NOTES | 12. SPONSORING MILITARY ACTIVITY |
|---|---|
| | Advanced Research Projects Agency<br>1400 Wilson Boulevard<br>Arlington, Virginia 22209 |

13. ABSTRACT

A strategy is outlined for acoustic aspects of speech recognition, whereby prosodic features are used to detect boundaries between phrases, then stressed syllables are located within each constituent, and a partial distinctive features analysis is done within stressed syllables. Facilities have been implemented for linear prediction, formant tracking, and extraction of fundamental frequency and speech energy contours. A program has been implemented which detects 90% of all boundaries between major syntactic constituents from fall-rise valleys in fundamental frequency contours. Improvements are planned which will reduce the false alarm rate and also incorporate energy cues to boundary positions. Studies of stress patterns in the Rainbow Script, read by six talkers, are being performed as a guide to developing techniques for automatic stressed-syllable location. A syntactic analysis and application of stress rules will yield theoretical preditions of stress. Acoustic analysis has yielded contours of fundamental frequency, energy, spectral data, and formant values throughout the spoken texts. Listeners' judgments are obtained, for each syllable, as to whether it is stressed, unstressed, or reduced. Incomplete results show considerable agreement (from talker to talker, listener to listener, and repetition to repetition) about stress judgments. Perceptual, acoustic, and linguistic results must be compared and interrelated. Speech texts are being designed to isolate factors affecting success in speech recognition, and cooperation with other ARPA contractors is directed toward integrating prosodies into total speech understanding systems.

DD FORM 1 NOV 65 1473



Unclassified
Security Classification

| 14 KEY WORDS | LINK A | | LINK B | | LINK C | |
|---|---|---|---|---|---|---|
| | ROLE | WT | ROLE | WT | ROLE | WT |
| Speech Recognition | | | | | | |
| Speech Analysis | | | | | | |
| Linguistic Stress | | | | | | |
| Prosodies | | | | | | |
| Prosodic Features Extraction | | | | | | |
| Syntactic Boundary Detection | | | | | | |
| Distinctive Features Estimation | | | | | | |
| Syntactic Analysis | | | | | | |
| Syntactic Parsing | | | | | | |

ik

SPERRY RAND

UNIVAC

# PROSODIC AIDS TO SPEECH RECOGNITION

by

Wayne A. Lea
Mark F. Medress
Toby E. Skinner

Defense Systems Division
St. Paul, Minnesota
(612-456-2430)

Semiannual Technical Report Submitted To:

Advanced Research Projects Agency
1400 Wilson Boulevard
Arlington, Virginia 22209

Attention: Dr. L. G. Roberts

Report No. PX 7940

1 October 1972

This research was supported by the Advanced Research Projects Agency of the Department of Defense under Contract No. DAHC15-72-C-0138, ARPA Order No. 2010, Program Code No. 90536. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the Advanced Research Projects Agency or the U. S. Government.

1c

## SUMMARY

Computers that understand speech are expected to facilitate natural man-machine interaction, but the problems involved demand the attention of several disciplines including linguistics, computer systems design, perception theory, speech research, and engineering. Linguistic and perceptual arguments, in particular, suggest that devices which recognize speech will have to make use of grammatical structure ("syntax") in early stages of the recognition procedures. This can be accomplished, in part, by using certain acoustic features, called the prosodic features, to segment the speech into grammatical phrases, and to identify those syllables that are given prominence, or stress, in the sentence structure.

Prosodic features (which include the durations of vowels and consonants, and time-varying measures of the rate of vibration of the talker's vocal cords and the energy in the speech) also provide some cues to the grammatical categories of phrases, the semantic associations between phrases, and the positions of reliable data for determining the sound sequence and word content of the speech.

Research described in this report is concerned with developing methods for detecting stressed syllables and the boundaries between grammatical phrases, using prosodic features. A speech recognition strategy is outlined which begins with detecting boundaries between phrases, by finding positions where the acoustic data show a substantial decrease in the rate of vocal cord vibration (that is, the "voice fundamental frequency"), followed by an increase in fundamental frequency. Once the connected speech is thus segmented into phrases, the strategy calls for locating the stressed syllables in each phrase. Then, analysis of reliable distinguishing features of the vowels and consonants within the stressed syllables is attempted. Speech sounds are expected to be more clearly articulated and easier to distinguish in stressed syllables, than in unstressed or slurred syllables, where articulation (and consequent acoustic data) is not as precise or consistent from talker to talker or time to time.

All the facilities for implementing this general strategy have not been implemented. A computer program for detecting boundaries between phrases succeeds

for about 80 to 90% of the expected boundaries, but also gives some "false" boundaries that are not apparently associated with syntactic structure. Improvements of this program are being planned, using energy contours and some detailed refinements in the test for substantial changes in voice fundamental frequency.

Facilities have been implemented for analyzing the acoustic speech signal to determine the resonances of the vocal tract (formants) that may indicate the vowels or consonants intended by the talker. These formant-monitoring facilities use frequency spectra derived from a recent analysis technique call linear prediction, which extracts the transfer function of the talker's vocal tract, de-emphasizing the harmonic structure of the vocal cord excitation source. Plots of the frequency of formants versus time are obtained from the smoothed frequency spectra provided by linear predictor analysis.

A program has been implemented for obtaining voice fundamental frequency and two measures of energy in the speech wave. These may be plotted versus time, or displayed on an interactive cathode ray tube terminal of the computer system.

A procedure for locating stressed syllables would represent a major component in the strategy for analyzing distinguishing features of speech sounds in stressed syllables within a constituent. Such a procedure for locating stressed syllables cannot be devised, however, until more is known about the acoustic features that mark presence of stressed syllables in connected speech.

Experiments are being performed to study stress patterns in a portion of a short text called the "Rainbow Passage." This text has been used extensively in studies of prosodic patterns in speech, and has the advantage of being a well-known semantically-connected text of declarative sentences, with a variety of grammatical phrase structures. The experiments with the Rainbow Script are designed to interrelate theoretical linguistic predictions, actual perceptual judgments, and acoustic data about stressed, unstressed and reduced (i.e., incompletely articulated) syllables in the connected speech. A grammatical analysis and use of published English stress rules will be done to provide theoretical predictions about stress patterns. An acoustic analysis has been performed

on the speech of six talkers reading the Rainbow Script, yielding contours of fundamental frequency, energy, spectral data, and formant values versus time. This acoustic data must yet be analyzed, to determine which acoustic features correlate well with predicted stress patterns and with perceptual judgments of stress.

When listeners heard clauses or sentences in the Rainbow Script repeated at will (by rewinding and replaying a tape), they were able to distinguish individual stressed, unstressed, and reduced syllables. Results differed little from talker to talker, or from listener to listener. Most differences that did occur indicated a difficulty in clearly distinguishing between unstressed and reduced syllables. One representative listener repeated the test several times, and demonstrated general repeatability of results. In one repetition, a computer was used in digitizing, storage, and replay, in place of the usual tape-rewind method. Under these conditions, the listener believed he could detect two levels of stresssed syllables ("highly stressed" and "lesser stressed"), besides unstressed and reduced levels.

About half the syllables were judged as stressed, while somewhat fewer were judged as reduced, and fewer than one quarter were judged as unstressed.

Further perception tests are yet to be made, with other listeners repeating the test, marking all syllables, and with repetitions to test consistency.

Studies of the relationships among the acoustic, perceptual, and linguistic data must be performed. Further tests are also being planned, using speech texts which are now being designed to isolate and study individual effects of position in the sentence, grammatical phrase structure, semantic structure, and phonetic content. Univac will also be evaluating speech data recorded by contractors who are building speech understanding systems for ARPA. This is one of many scheduled activities designed to integrate prosodic information into other programs on total speech understanding systems.

# TABLE OF CONTENTS

Page

# 1. INTRODUCTION

This is a report on work currently in progress in the Univac Speech Communications Group, under contract with the Advanced Research Projects Agency (ARPA). As a part of ARPA's total program in research on speech understanding systems, the research reported herein is concerned with extracting reliable prosodic and distinctive features information from the acoustic waveform of connected speech (sentences and discourses). Studies are being concentrated on problems of detecting stressed syllables and syntactic boundaries.

The traditional model of speech recognition has assumed that, by tracking the right "information-carrying" parameters, and using any of several phonemic-segment classification techniques, one could determine phonemic strings corresponding to those intended by the talker. Then, the phonemic strings may be applied to higher-level linguistic analyses to determine words, phrases, and utterance meanings.

At Univac, work on automatic speech recognition (ASR) has progressed along a different approach. The viewpoint is that versatile speech recognition will proceed by making use of reliable information in the acoustic data in combination with early use of linguistic regularities. As will be outlined in this report, recognition is to be accomplished by using prosodically-detected stress patterns and syntactic structure in aiding a partial distinctive-feature-estimation procedure. Prosodically-detected syntactic structure will also be used to aid syntactic parsers and semantic processors.

Prosodic cues to sentence structure, and prosodic aids to the location of reliable acoustic phonetic information, have been given little or no attention in previous speech recognition efforts. The strong motivations for the use of prosodic patterns in speech recognition procedures will thus be presented in some detail in section 2. Versatile facilities for extracting prosodic features, spectral data, and formants, and a program for detecting boundaries between syntactic phrases (constituents), will be described in section 3. Initial experiments to be described in section 4 are being conducted to: determine the acoustic correlates of stress and constituent boundaries; determine listeners'

abilities to perceive stressed, unstressed, and reduced syllables; and derive theoretical linguistic predictions about stress patterns and syntactic boundaries in English sentences. These experiments are being performed on a well-known connected text called the "Rainbow Script" (Fairbanks, 1940), but further studies will be conducted later on texts specifically designed to isolate interfering factors in prosodic-phonetic-syntactic interaction in connected speech. In section 5, efforts to design good speech texts are described, along with efforts to integrate prosodic studies with research on other aspects of total speech understanding systems.

## 2. MOTIVATION FOR PROSODIC AIDS TO SPEECH RECOGNITION

Speech is man's most natural, universal, and familiar form of communication. It has many advantages which have been shown to apply to man-machine interaction as well as to human communication (Lea, 1968; 1970). Among the most difficult problems involved in speech communication between man and computer is the computer recognition of speech. Speech recognition might be defined as the process of transforming the continuous acoustic speech signal into discrete representations which may be assigned proper meanings, and which, when comprehended in a total speech understanding system, may be used to affect responsive behavior.

### 2.1 Arguments Favoring Prosodic Cues to Syntactic Structure

Early work on speech recognition was concerned with pattern matching on isolated words, achieved by direct comparison of input spectral data with stored spectral patterns (or "templates") obtained from previous processing of the words in the vocabulary. Later work acknowledged the phoneme as a recognizable segment. Word recognition was to be done by recognizing phoneme strings as constituting words. Probabilities of phoneme sequences were expected to help increase the accuracy of word recognition algorithms.

As interest in recognition of continuous speech developed, the general problem of speech recognition was regarded as being composed of two parts: "a primary recognition based solely on the sound shapes of the acoustic signal; a secondary recognition of the linguistic (grammatical and syntactic) content based on the (presumably phonemic) output of the primary recognition level" (Lindgren, 1965). The prevalent hope was that one could segment the acoustic stream into moderate-sized discrete atoms (phonemes, diphones, or such), which could be independently recognized.

#### 2.1.1 Linguistic Arguments

There are two faulty assumptions implicit in such a hope. One, often called the linearity condition, asserts that there should be a distinguishable segment in the speech wave for each abstract (phonemic) segment, and if abstract segment A precedes abstract segment B in the abstract linguistic string, then the time

segment associated with A must precede that for B. The other assumption, called the invariance condition, asserts that all the distinguishing features of an abstract segment must be present (within the segment's time-stretch) for each occurrence of that segment, while the set of all such features values should not occur for other abstract segments. As Chomsky and Miller (1963, p. 311) noted, "If both the invariance and linearity conditions were met, the task of building machines capable of recognizing the various phonemes in normal human speech would be greatly simplified". However, violations of invariance and linearity abound. For example, the distinction between the words ladder and latter (phonemically, /læ d ɚ/ vs /læ t ɚ/), which is phonemically in the third segment, physically occurs often in the lengthening of the second phonetic stretch of sound (phonetically, læ·Dɚ vs læ Dɚ). This violates both the linearity and invariance conditions. Any coarticulation process, or context dependency, whereby a (nearby or distant) phoneme causes changes in the distinguishing acoustic properties of a given phoneme, would similarly violate the two conditions.

Another example of violations of linearity and invariance concerns an additional way in which voicing of consonants is marked in the acoustic data. Voicing of some voiced consonants is not always evidenced by the expected continuous periodic vibration of the vocal cords throughout the consonant. Both voiced and unvoiced consonants may have initial "voiced" portions of their closure period during which the vocal cords vibrate periodically, followed by "unvoiced" portions during which periodic vibration does not occur. When such discontinuities in vocal cord vibration occur, voicing is not determined by features within the closure period associated with the consonant. Secondary features outside the time stretch of the consonantal closure, such as the initial rate of change of the fundamental frequency of vocal cord vibration within the following vowel, must be used to establish phonemic voicing (cf. Stevens, 1971; Lea, 1972a, Ch. 4). As noted by the example above, voiced consonants also cause a lengthening of the duration of the preceding vowel. Consequently, a speaker has the option of producing voiced stops and fricatives without actually vibrating his vocal cords continuously, provided he increases the duration of the preceding vowel, or otherwise supplies cues somewhere within the utterance as to the [+voiced] state of the consonant.

Stress and intonation are other linguistic factors that show marked physical effects in one segment (vowel or syllable) due to surrounding segments. Violations of linearity and invariance occur so frequently that the linguists have written quite general phonological rules (such as the one for lengthening of vowels before voiced consonants, or the context-dependent stress rules) to capture such generalizations. Because of the structural redundancy provided by the listener's linguistic knowledge, a speaker does not have to encode into the acoustic waveform all of the features describing an utterance, and the features that he does choose to encode can vary from one repetition of a given utterance to the next. A listener will be able to fill out the distinctive features matrix for a word he has heard, knowing only some of the matrix elements and using his knowledge of the structure of the language. For example, in the feature matrix representation of the single morpheme word "slump" shown in Figure 1, a total of 39 matrix elements is used to specify the five phonemes. However, if full use is made of the structure of English, the 24 unshaded matrix elements can be derived from knowledge of only the 15 that are shaded in. To do so in this example, one would utilize the facts that /s/ is the only sound that can precede an initial /l/, and that if a single morpheme word has a final consonant cluster beginning with a nasal, the following consonant must share place of articulation. Of course, the number of features necessary for identification could be less than 15 if one was dealing with a restricted and limited lexicon, and considerably less if the word occurred in a sentence or phrase that further limited the number of choices available to a listener.

The combination of linguistic and lexical structure, and multiple cues for some features, allows a speaker to thus be imprecise and inconsistent in his production and still be clearly understood. The net result is that in addition to the fact that the encoding of phonemic and prosodic information into the acoustic waveform is a complex one involving overlapping in time and environmental dependence, the encoding itself is often performed incompletely and with considerable variability. Indeed, in some utterances, whole phonemes or syllables may be "missing" from the pronunciation. A speech recognition system based on acoustic manifestation of all phonemes or all distinctive features would thus frequently fail.

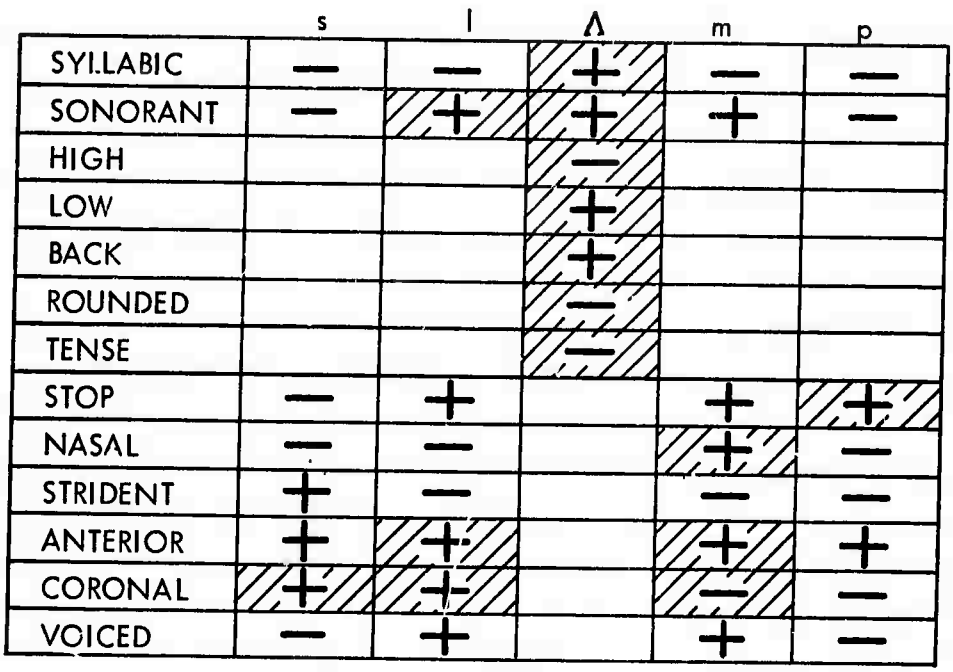



Figure 1. Distinctive features matrix for "slump." Features that play no role in describing a particular phoneme are left blank. The significance of the shading is explained in section 2.1.1.

These and other arguments (cf. Chomsky and Miller, 1963; Lea, 1972a; IN PRESS; Medress, 1972) against invariance and linearity in small speech units dispel any notions that recognition based on simple concatenation of categorized time segments can be completely successful. Phonemic context has thus been recognized as necessary to fill in acoustically "unspecified" distinctive features in the representation of received utterances, and to even fill gaps for "missing" phonemes in some pronunciations of connected speech.

However, linguists also argue (cf. reviews by Lea, 1972a, b) that phonemic recognition, or distinctive features estimation, cannot even by accomplished with the use of phonemic context and known phonetic redundancies. They argue that "in general, the perceiver of speech should utilize syntactic cues in determining the phonemic representation of an utterance" (Chomsky and Miller, 1963, p. 314; emphasis added). Chomsky and Halle (1968, p. 31) for example, have developed a detailed set of phonological rules for (phonetic) stress assignment, along with vowel reduction rules and other phonological rules, which depend explicitly on the word categories and phrase structure of utterances. Such rules are assumed to be used by the speech perception system in relating phonetic data to linguistic structure.

### 2.1.2 Perceptual Arguments

Psychologist George Miller (1962) also has sharply criticized the view that speech recognition should be achieved by first deciding what phonetic segments have occurred, then determining what phonemes and morphemes were involved based on the lower-level phonetic decisions, and so on up to larger units and higher linguistic levels. He gives several reasons for doubting that people naturally operate that way, (see also Chomsky, 1964, pp. 106-114, and Flanagan, 1965, pp. 236-8, for other arguments). If we are to assume that a recognizer will exhibit behavior similar to that of the human perceiver, we may also doubt the value of such a model for artificial recognizers. Miller argues (1962, p. 81) as follows:

> Phenomenologically, it seems that the larger, more meaningful decisions are made first, and we pursue the details only so far as they are necessary to serve our immediate purposes... If the small details of input are discriminated first, how is it possible

> to take advantage of the redundancy of the message?... we must regard the decisions reached at the lower levels as tentative and subject to revision pending the outcome of decisions made at some higher, more molar level. Once this tentative character is admitted, of course, it becomes necessary to continue storing the original input until the molar decisions have been reached. However, if complete storage is necessary even after the lower-level decisions have been tentatively reached why bother to make the lower decisions first?

Arguing from reaction time studies, Miller asserted that we just do not have enough time to make all phonemic decisions at the rate at which phonemes occur in speech. He estimated that about one categorical decision per second might occur in ordinary listening, and concluded that: "If we accept this as a rough estimate, it suggests that the phrase -- usually about two or three words at a time -- is probably the natural decision unit for speech" (Miller, 1962, p. 81). Miller reported some experimental results that supported his conjectures.

Other perception studies have confirmed the use of phrases as units, at whose boundaries decisions appear to be made. Johnson (1965) showed that the probability of an error in remembering the next word in a sentence increased significantly at phrase-structure boundaries, thus indicating that sentences are remembered phrase-by-phrase. Similar studies had previously been made showing that probabilities of error in predicting phonemes increased markedly at word and phrase boundaries. Several other studies showed that clicks superimposed on speech were perceived as occurring near certain major deep-structure syntactic boundaries within the sentence, regardless of actual timing of the clicks within the speech continuum (cf. review by Gleitman and Gleitman, 1970).

It has been suggested that the perceiver waits until the end of such phrase units ("constituents") before making decisions as to the sound structure content of the large unit. The timing of a click is lost since decisions as to its phonetic significance are delayed until the end of the constituent, at which time its relationship to the rest of the sound sequence is found to be nil, and its relationship to the time structure of the recognized large unit cannot be established. Fodor and Garrett (1966) in particular, noted that constituents dom-

inated by a sentence node (in deep structure) yielded particularly regular click displacement to their boundaries. The title of a recent paper by Bever, Lackner, and Kirk (1969), nicely summarizes the usual interpretation of these studies: "The Underlying Structures of Sentences Are the Primary Units of Immediate Speech Processing".

Such results suggest that people generally make slow, infrequent decisions about relatively large units of speech, rather than many fast decisions about small units. The results do not, of course, imply that phrases are the only units involved in perception (cf. Haggard, 1967).

Miller (1962) observed that the use of large units in the detection of the message in speech is not surprising in the light of studies in coding theory. Messages can be more efficiently encoded, and error-correcting information can be introduced, if a long string, or long segment, is stored and encoded (and later decoded) as a unit.

It is also interesting that children, as primitive speech-recognizing systems, learn intonational cues to phrase structure and sentence type before acquiring any competence with the specific phonemics of their language community (Lieberman, 1967a; Lewis, 1936; Leopold, 1953). On another hand, Grimes (1969) has shown that field linguists trying to perceive the structure in a new language benefit from early use of large-unit segmentation into "breath groups" and rhythmic (sense group or phrase) units.

If even phonetic and phonemic decisions involve syntactic information, as suggested in section 2.1.1 above, then Miller would seem to be right in suggesting that advantageous use of the redundancy of language and speedy perception require making higher-level decisions about large decision units (such as phrases) before firm decisions are made about lower-level phonemic units. One might conclude that there is overwhelming linguistic and perceptual evidence suggesting the need for early introduction of syntactic hypotheses in recognition schemes.

2.1.3 Prosodic Cues to the Presence of Large Linguistic Units

These arguments suggest a somewhat novel theory of speech recognition, using syntax in phonemic decisions. Speech perception then involves making use of certain expectations and received cues to determine the syntactic structure (and semantic content) of an utterance. Given a hypothesis as to the surface syntactic structure, the perceiver uses phonological principles to determine a phonetic shape. The hypothesis will be accepted if its associated acoustic phonetic shape isn't too radically different from the acoustic input (Chomsky and Halle, 1968).

How might one make the preliminary syntactic hypotheses called for in the early stage of recognition schemes, without depending upon a preliminary segmental (phonemic) analysis? The listener must presumably be using some cues in the acoustic signal to guide his hypothesis-making. What acoustic cues or features might be used? Obviously they must be features which extend throughout the large units of syntactic structure, or they must be localized features that mark unit boundaries, centers of units, or some such critical points in the structure. The boundary-marking features, often identified as "junctures" (Peterson, 1963; Delattre, 1965), disjuncture (Lieberman, 1967), or deliminative and culminative elements (Trubetskoy, 1939, p. 27), signal the boundaries between two units and indicate how many 'units' are contained in a particular sentence or other extended utterance. Other features extending over a large unit may provide a distinctive function which identifies the class of a unit and distinguishes that unit's category from other possible structural categories.

Among the features that are known to provide deliminative and distinctive markings of syntactic units are the prosodic features. (Other features, such as the allophonic variations between word-initial aspirated stops and word-final unaspirated stops, would also be relevant.)

Prosodic features that have long been recognized as indicators of English constituent structure are voice fundamental frequency (abbreviated as $F_0$), speech intensity, and the relative durations of phonetic segments. For example, vowel and consonant durations are known to increase just before pauses between

syntactic units (Allen, 1968; Barnwell, 1971; Mattingly, 1966). Lieberman (1967, pp. 152-3) showed that, for some lower-level syntactic disjunctures, such as the distinction between "light housekeeper" and "lighthouse keeper", the time interval between vowel centers is a reliable cue for disjuncture positions. Prosodic features also closely relate to stress patterns, which in turn are closely associated with the syntactic bracketing and syntactic categories in sentences (Chomsky and Halle, 1968; Lea, 1972a, Ch. 6).

## 2.2 Prosodic Cues to Boundaries Between Phrases

For decades, linguists have claimed that intonation (that is, the perceived variations in the pitch of the talker) indicates the immediate constituent structure (i.e., "surface structure") of English sentences (Jones, 1909; 1932; Pike, 1945; Hultzen, 1957; 1959; Wells, 1947). Trager and Smith, whose pitch and stress "levels" (1951) are widely used, claimed that monitoring voice fundamental frequency ($F_o$) makes it possible to have "solidly established objective procedures" for "the recognition of immediate constituents and parts of speech syntax" (1951, p. 77). Yet, they did not say exactly how to use intonation for structural analysis, and, until recently, no such "objective procedures" had been publicized. Gleason (1961, p. 169) also considered intonation and stress as "the dominant elements in the syntax-signaling system". Study of metrical patterns in English verse also indicate strong markings of syntactic boundaries by the prosodic features (Keyser, 1969).

Transformational linguists have also recognized this syntax-signaling role of intonation. Lieberman (1967, p. 314) asserted that:

> "Intonation has a central role in the transformational recognition routines that the listener must use for syntactic analysis. Intonation provides acoustic cues that segment the speech signal into linguistic units suitable for syntactic analysis."

Stockwell (1960) noted that, "There is a good deal of evidence . . . that intonation patterns are the absolutely minimal differentiators of numerous utterance tokens." That is, intonation helps disambiguate structurally ambiguous utterances by indicating their bracketing into syntactic units. Bierwisch (1965) demonstrated that it is possible to generate an intonation contour (for

a German sentence), if only the surface syntactic tree and related syntactic information is provided.

While there has been widespread agreement about intonation marking some boundaries between sentence parts, there has been considerable dispute about how this is accomplished (see review by Lea, 1972a, section 1.4). Some issues are the following: (1) What features of the voice fundamental frequency contours mark boundaries between subunits of sentences? (2) Which sentence portions (major grammatical constituents, clauses, all syntactically bracketed units, or arbitrary sequences of words) are actually demarcated by intonational features? (3) Are syntactic units demarcated by intonation patterns in all utterances, or only when the talker is explicitly trying to clarify the structure of structurally ambiguous utterances?

One of the weakest hypotheses about intonational cues to sentence structure (Armstrong and Ward, 1926) is that sentences may (but need not necessarily) be divisible into parts by intonation contours associated with any arbitrary (but fairly long) sequences of syllables or words. The units need not be syntactic constituents, and indeed the individual talker may divide (or not divide) a sentence differently from time to time, and different talkers may divide utterances differently. At the other extreme, all sentences are assumed to be divisible into syntactic units by intonational (or other prosodic) cues that always occur at unit boundaries (Trager and Smith, 1951; Wells, 1947).

Invariance applied to such prosodic aspects of language would imply that a syntactic boundary always has an associated acoustic (or phonetic) boundary marker manifested, and only when the syntactic boundary occurs will that acoustic marker appear (Trager and Smith, 1951, p. 51). Linearity would imply that a boundary between two syntactic units would be manifested by acoustic features at the time stretch ('pause' or such) after the time stretch associated with the last phoneme of the earlier constituent and before the time stretch associated with the phonemes of the later constituent.

Malmberg (1963, p. 69) implicitly rejected the linearity condition for structural boundaries. He broke up utterances into "measures" on the basis of

perceived intonation, yielding such divisions as the following:

The boys are —— playing in the —— street.
ðə bɔɪz ɚ —— pleɪɪŋ in ðə —— strit

Each measure or group has an "accented" (stressed) syllable and zero or more unstressed syllables. The breaks he shows occur just before the stressed syllables, and not necessarily at the points in the phonemic string where structural boundaries occur. This break-down into groups is exactly what is obtained from the automatic analysis of fundamental frequency contours, to be described in Section 3.4. That is, strict linearity must be rejected if one is to succeed in finding acoustic cues to the syntactic breaks.

Recently, Robert Scholes (1971, pp. 50-73) investigated whether fundamental frequency, peak amplitude in syllabic nuclei, or inter-vowel intervals provided the best cues as to whether a syntactic (subject-predicate) boundary occurred. Listeners were presented with three contiguous words extracted from one of eight sentences (read by any one of ten talkers) and were asked to judge whether a subject-predicate boundary occurred between the first and second, or between the second and third, words. About 84% of all sugject-predicate boundaries were perceived in the correct positions by the panel of listeners. For these perceived boundaries, he found that the time interval between vowels did not usually correlate well with whether or not a subject-predicate boundary was between the two syllables. However, the syllable that is perceived as phrase-final was more intense (higher in VU reading) than the preceding or following (non-phrase-final) syllables. No strong generalizations could be made from Scholes' study of fundamental frequency ($F_0$) contours, primarily because he only investigated whether $F_0$ increased or decreased within a syllable, not how $F_0$ values in one syllable related to those in other syllables. Other studies (Lea, 1971, 1972a, 1972b) show good correspondence between $F_0$ contours and boundaries between syntactic units.

As pointed out in a recent study (Lea, 1972a, section 1.4), most attempts to find acoustic cues for syntactic boundaries have involved some of the most questionable syntactic boundaries possible, including the subject-predicate boundary of a sentence, and such small-unit distinctions as <u>light</u> <u>housekeeper</u> versus <u>lighthouse</u> <u>keeper</u>. Studies to be reported on in sections 3.4 and 4 are designed

to test for prosodic cues to constituent boundaries in a variety of positions in syntactic structure.

## 2.3 Prosodic Cues to Stress Patterns and Categories

Besides allowing segmentation of sentences into syntactic units, prosodic features can also provide some cues to the categories of syntactic units (such as sentences, nuclear noun phrases, compound nouns, etc.). This would be accomplished primarily through using prosodic features to determine stress patterns, which are known to associate closely with syntactic bracketing and syntactic categories.

Since fundamental frequency ($F_o$) and intensity both tend to be higher, and phonetic durations tend to be longer, for stressed than for unstressed syllables, monitoring $F_o$ contours and intensity contours might determine the relative stress levels throughout the utterance (Lea, 1972a; IN PRESS, p. 200; Hughes, Li, and Snow, 1972; Medress, Skinner, and Anderson, 1971). The following conjecture is then suggested:

> "If one could, by tracking acoustic features such as voice fundamental frequency (pitch), average speech power (intensity), and phonetic durations, determine the stress pattern(s) of an utterance, and if such stress patterns could predict vital aspects of surface syntactic structure, then one might be able to use such prosodic information to automatically guess at surface syntactic structures." (Lea, IN PRESS, p. 200)

To test such an idea, one must have: (1) an adequate procedure for determining stress patterns from acoustic features; and (2) an adequate set of rules for predicting syntactic structure given that the stress pattern can be established.

Various schemes for determining stress from acoustic features havo been or are being investigated (Fry, 1955; Hughes, Li, and Snow, 1972; Lieberman, 1960; 1967a; Medress, Skinner, and Anderson, 1971). Further studies of the acoustic correlates of stress in connected speech are needed. In fact, such studies of stress constitute a major portion of Univac's present and proposed efforts for ARPA. Experiments with the Rainbow script, to be described in section 4, are aimed at determining (among other things) the correlation between parameters of $F_o$ and energy contours and the perceived and linguistically-predicted

stress patterns. Further studies will be proposed for determining acoustic correlates of the stress patterns in a variety of specially designed sentences, so that interfering effects of sentence type, syntactic categories, phonemic content, and the like can be independently isolated.

Even when one finds the most reliable cues to stress patterns, his job of syntax recognition is far from done. He must relate stress back to abstract syntactic structures. However, it is not easy to relate stress patterns back to syntactic structures, for the purpose of establishing the categories of constituents and the sentence bracketing. A few informal rules relating stress and syntactic categories have been known for some time. For example, it is well known among phonologists that monosyllabic function words such as articles, prepositions, anaphoric pronouns, and conjunctions are characterized as unstressed (or "weakly stressed") (Halle and Keyser, 1971, p. 9). "Substantives" like nouns, and most verbs and adjectives, are often stressed, particularly if they are polysyllabic. Thus, if a word or syllable is found to be stressed, it is more likely to be a noun, verb, or adjective, (Chomsky, 1965, Ch. 2) than a function word (sometimes called a "grammatical formative"; cf. Chomsky, 1965, Ch. 2).

Chomsky and Halle (1968) provided explicit rules for relating syntactic categories (and phonemic content of words) to stress patterns. Several revisions to their rules have been suggested (Halle and Keyser, 1971; Vanderslice, 1969), but certain essential features are common to all such formulations of English stress rules. Certain lexical stress rules (which depend upon the phonemic content and category of a word) dictate which syllables in polysyllabic words are stressed. When words are grouped into a constituent, two major stress rules apply. One is the Nuclear Stress Rule (Chomsky and Halle, 1968; Halle and Keyser, 1971), which says that if several lexically-stressed syllables (vowels) occur in a constituent (called an "$\alpha$-constituent") not labelled as a noun (N), verb (V), or adjective (A), then the last of these stressed syllables gets the primary stress and the others are reduced in stress level with respect to it. The other major rule, called the Compound Stress Rule, says that for constituents labelled N, A, or V, the first of such lexically stressed syllables gets primary stress. The rules apply cyclically, starting from the smallest constituent within brackets and working out

to the whole sentence. From such rules, there is a strong tie between the categories of constituents (N, A, V versus $\alpha$) and the stress pattern. One might hope then that, knowing the stress pattern, he could work backwards to the syntactic bracketing and categories. In actual fact, these reverse (stress-to-syntax) rules would be quite complex, and no one-to-one map from stress to syntactic structure would be possible (Lea, 1972a, p. 163f). Still, an attempt could be made to define some such stress-to-syntax rules that may be appropriate for speech recognition procedures.

## 2.4 Prosodic Aids to Distinctive Features Estimation

Most of the work outlined in the previous section is concerned with recognizing syntactic structure from prosodic patterns, without the use of any segmental phonemic information. But, a speech understanding system must ultimately use distinctive features information, plus syntactic parsers and semantic processes, in the total effort in sentence recognition.

One way in which prosodic information, and resulting syntactic segmentation and stress pattern analyses, may be used in distinctive features estimation is as follows. At an early stage in recognition, one detects boundaries between major syntactic constituents from prosodic features. (A technique for doing such is described below, in section 3.4.) Then, the highest-stress syllable(s) within each constituent is (are) located, using reliable prosodic cues to stress. (Techniques for stressed syllable location, as planned at Univac, are outlined in section 4.) Some distinctive features are then to be estimated within these stressed syllables, since the consonants and vowels are expected to be easier to categorize in stressed syllables than in weakly stressed or reduced syllables (Hughes, Li, and Snow, 1972). Next, the partial distinctive features description is matched with generated or stored patterns for possible stressed syllables or words in the lexicon. Then a guess as to the word content of the constituent is made, based on the reliable feature information from the stressed syllable(s), plus other reliable data within the constituent (such as presence of unvoiced coronal strident fricatives, etc.; cf. Medress, 1969, 1972). If reliable decisions cannot be made based on such minimal feature information within the constituent, analyses are then applied to other words or syllables at lower stress values, and a guess based on the two or more moderately-stressed syllables is made. Iteration would continue until all syllables are analyzed, if necessary. Each

iterative guess as to constituent identity would be combined with those for other constituents in the sentence until a satisfactory set of hypotheses for all constituents yielded the grammatical, meaningful sentence.

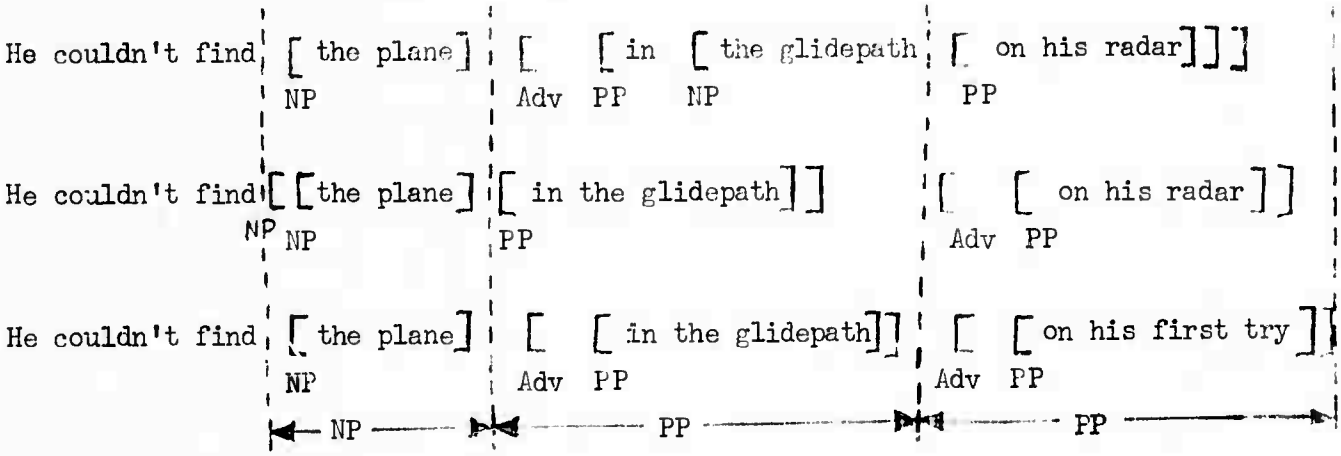
One assumption in this approach is that the phonetic dependencies across constituent boundaries will be considerably less than the interdependencies within a constituent. Then recognition of these substantially isolatable constituents could be attempted, presumably more reliably than context-independent phonemic segmentation and classification could be achieved. This would constitute an assumption that linearity and invariance conditions are essentially satisfied for constituents. Previous studies (Lea, 1972a) showed that full-clausal embedded sentences and matrix sentences were separated by long pauses. Certainly, phonetic dependencies might be expected to be less likely across such structural pauses, and one might hope that this will also be true for smaller, more manageable constituents.

The essential ingredient of this type of approach is that speech recognition involves using prosodic features to make early hypotheses about syntactic structure, which then can be used to guide distinctive features estimation processes. Ultimately, what is sought are prosodic cues to the phonological rules which have applied to surface syntactic structures to yield the observed acoustic data.

## 2.5 Prosodic Aids to Syntactic Parsing

The methods described in the previous section for guessing the word content of each constituent depend upon determination of aspects of syntactic structure before the terminals ("leaves") of a syntactic tree are determined (cf. Willems, 1972; Lea, 1972a). Syntactic parsers, on the other hand, usually address the problem of determining the labelled bracketing of a sentence, given the terminal string as input information. A major task is to establish how prosodically-determined syntactic structure may be used to aid syntactic parsers. How can one use the syntactic segmentation and stressed-syllable-location procedures to help disambiguate terminal strings? Part of the answer lies in determining specific problem areas in parsing that could be helped by knowledge of stress or boundary information.

For example, structures of the form Noun Phrase - Prepositional Phrase - Prepositional Phrase are said* to give particular difficulties to syntactic parsers, in that the associations between the phrases can be quite different from utterance to utterance. The following partially structured sentences illustrate a few cases where different structures and semantic associations have similar NP-PP-PP surface structures:




## 2.6 Other Uses of Prosodic Cues

There is some possibility of detecting aspects of semantic structure from prosodic patterns. It is known that emotion and some semantic distinctions (uncertainty, incompletion, doubt, etc.) affect intonation and other prosodies (Armstrong and Ward, 1926; Huttar, 1968). Also, grammatical relations such as coreference, contrast, antecedent-pronoun associations, etc., have been said by linguists to have regular effects on intonation (Cantrell, 1969). Recent rules for stress assignment (Bresnan, 1971, 1972) assume that relative stress levels (determined by the nuclear stress rule) are dictated by the embedded deep structures of sentences, which are applied through the iterative syntactic transformational cycle. Since deep structures are closely associated

*According to personal communication with Jerry Wolf, BBN.

with semantic interpretations of sentences, stress levels (and thus their acoustic correlates) might then be relatable to underlying semantic structures. These claims about semantic cues in prosodic patterns must be instrumentally investigated.

## 3. SYSTEMS FOR EXTRACTING PROSODIC AND DISTINCTIVE FEATURES

In section 2, the motivation was given for a program for applying prosodic features to the detection of stress and syntactic boundaries. To use prosodic information in cooperation with a partial distinctive features estimation procedure, facilities are required which extract prosodic features of fundamental frequency, speech energy, and timing and durational data, as well as spectral data and formant values versus time. In this section, we describe such facilities as implemented at Univac, Defense Systems Division (DSD), and their use in a computer program for detecting sentence boundaries, constituent boundaries, and other cues to syntactic structure. These feature-extraction and structure-detection facilities will be coupled with techniques of stressed-syllable location to provide acoustic guidelines to syntactic parsers, semantic processors, and procedures for identifying the lexical identity of distinctive features patterns.

The speech analyzing capabilities include methods for linear predictive analysis and formant tracking (section 3.2), prosodic features extraction (section 3.3), and syntactic constituent boundary detection (section 3.4). These analysis tools are operating within a versatile interactive speech research facility, to be described next.

### 3.1 Interactive Speech Research Facility

The Univac speech research facility is a highly flexible and interactive system designed especially for processing and studying speech. The speech facility is located in the Speech Communications Laboratory adjacent to the Univac DSD Engineering Computer Center. In addition to fabrication, testing, and storage facilities, the laboratory contains a 12' by 12' Industrial Acoustics Corporation sound isolation room. This room provides an extremely quiet environment for the speech research terminal. High quality audio tapes with no significant backgound noise can be made there for subsequent analysis, or speech may be entered directly into the computer for study.

A block diagram of the system is shown in Figure 2. With this facility speech can be appropriately filtered and digitized at up to 20 kHz and stored on a drum. It can then be played back over a speaker or displayed on a cath-

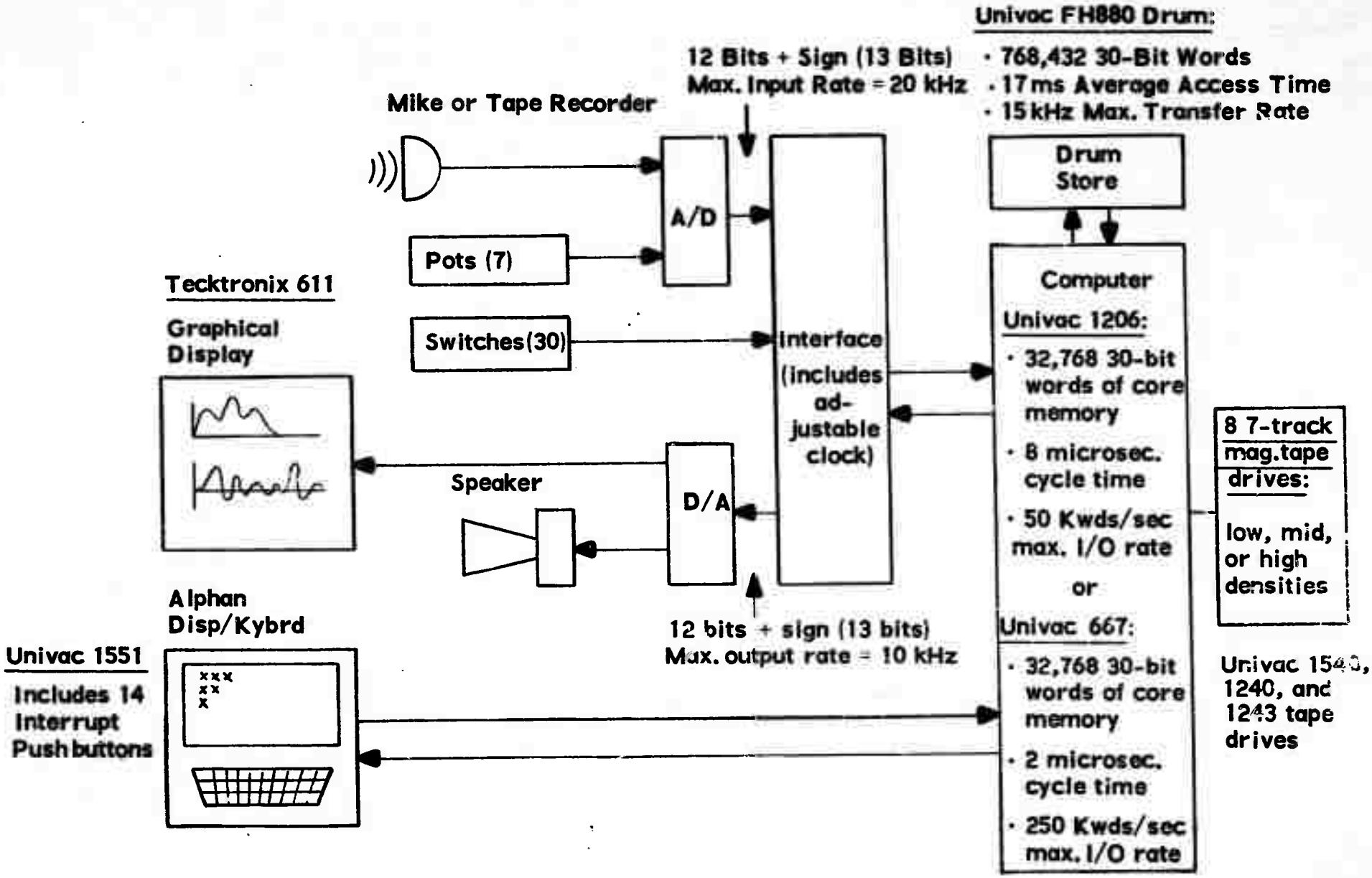



Figure 2. Block diagram of the Univac interactive speech research facility.

ode ray tube (CRT). The digitized acoustic waveform can also be processed by fast Fourier transform (FFT) or linear prediction programs to obtain short-time spectral patterns, and by other algorithms to generate fundamental frequency and energy contours, formant tracks, and other data of interest. All of the data can then be simultaneously displayed and examined on the CRT. Interactive control of the system is provided by toggle switches, push buttons, potentiometers and an alphanumeric display and keyboard.

In order to study long utterances and inter-sentence effects, the research facility can accept and process up to 12 seconds of speech at one time. The interactive display can be used to simultaneously examine the time waveform, smoothed spectra, and up to 20 time functions, including formant tracks, voicing and fundamental frequency contours, and various frequency delimited energy functions for a full 12 seconds of speech.

A digital tape storage facility has been developed so that a local data base can be built up during the course of speech studies. This facility has been designed to provide fast and easy access to previously processed speech data for reexamination and further processing. It will complement and be compatible with the Lincoln Speech Data Facility.

Finally, steps are being taken to connect the speech research facility to the ARPA network through a Very Distant Host Interface and a 50 kilobit line. Software and hardware design should be completed by mid-November. It is anticipated that the network connection can be implemented in February, 1973.

## 3.2 Linear Prediction and Formant Tracking

The technique of speech analysis by linear prediction has been implemented on the speech research facility.* This analysis technique produces very high quality smoothed spectra. A formant tracking program similar to one of Shafer and Rabiner (1970) has been developed utilizing the smoothed spectra obtained from linear prediction, thus making formant information now available for study and use.

---

*In the Univac implementation of linear prediction, the assistance of John Makhoul (Makhoul and Wolf, 1972) has been appreciated.

In the experiments with the Rainbow Script (described in section 4), the acoustic waveform was first prefiltered to 4782 Hz with a seventh order elliptic function (Cauer) low pass filter provided by Lincoln Laboratory. The following parameter values were then used in the analysis: 10 kHz sampling rate (and thus a 5 kHz frequency analyzing bandwidth), 25.6 msec Hamming weighted spectral analysis window with a 10 msec advance, and 12 predictor coefficients. In addition, the speech signal was processed without pre-emphasis. This permits the evaluation of the linear predictor normalized error as a potential voicing function. It has been shown (Makhoul and Wolf, 1972) that the normalized error is not a good indicator of voicing if the speech has been pre-emphasized.

### 3.3 Prosodic Features Extraction

Fundamental frequency, energy, quality, and duration are useful prosodic features (Lea 1972a; Medress and Skinner 1972; Medress, Skinner, and Anderson, 1971). In the Rainbow Script experiments (discussed in section 4), various tabular and graphical time functions were extracted which are indicative of these prosodic features.

By autocorrelating the center-clipped acoustic time waveform (Sondhi, 1968), a fundamental frequency measure was obtained every 10 msec for a 51.2 msec time segment over a range of 70 to 400 Hz. Fundamental frequency in Hertz was also converted to eighth-tones, yielding a log frequency scale for relative measure. Alternative methods for fundamental frequency determination (cepstral analysis, Noll, 1967; linear prediction, Makhoul, 1972) have been implimented on the research system but, at present, do not perform as accurately as autocorrelating the center-clipped time waveform.

Various frequency-dependent energy functions were computed as part of the Rainbow Script experiments. One time function which reflects total energy in a 25.6 msec window every 10 msec was computed from the sum of the squares of the time waveform values (Blackman and Tukey, 1958). (This sum is the first autocorrelation term, an intermediate parameter of linear prediction spectral analysis.) Other energy measures obtained in the frequency domain include: (a) a low frequency, sonorant energy function, computed by summing the squares of the smoothed spectral magnitudes from 0 to 3000 Hz and then converting the sum to

dB, and (b) a broadband energy function, computed similarly by summing the squares of the smoothed spectral magnitudes from 0 to 5000 Hz and converting the sum to dB.

Other outputs from the Rainbow Script experiments are hoped to be useful as quality and durational indicators. These include digital spectrograms and formant tracks (both tabular and graphical), from which vowel reduction can be estimated. In addition, correlation (Hogg and Craig, 1965) and spectral derivative (Medress, 1969) time functions were computed. These functions indicate some vowel boundaries (and thus some phonetic durations). For example, at a vowel-obstruent boundary, a definite peak will occur in the spectral derivative function and a prominent valley will appear in the correlation function.

Figure 3 is a typical graph of total energy in dB, (as computed in the time domain) and fundamental frequency (in eighth tones) for speaker ASH reciting "they act like a prism", as extracted from his reading of the connected text of the Rainbow Script. A value is recorded for each function every 10 msec from time 3160 to 4880 msec. Energy is shown by the symbol "B" and fundamental frequency is indicated by the symbol "0". The tabular data at the top of the graph are fundamental frequency in Hertz, broadband energy in dB and fundamental frequency in tones. For example, at time 3750 msec, the energy graph is at 68 dB, while the fundamental frequency function is at 71 tones. Note from the tabular data at time 3750 that 71 tones corresponds to 192 Hertz.

Figure 4 is a photograph of the interactive graphical display. With appropriate potentiometer control, that portion of the acoustic time waveform which corresponds to the data of Figure 3 ("they act like a prism", speaker ASH) has been selected for display and is shown at the bottom. The number at the ordinate shows the time waveform display to begin at 3154.7 msec and include the time data to 4879.9 msec, the time indicated at the base of the time waveform cursor. At the top of the display is a spectral plot (relative amplitude in dB versus frequency) of a vowel portion of the utterance. The particular spectral frame that is displayed is selected with a potentiometer, and thus the short time spectral pattern can be examined throughout the entire utterance. Another potentiometer controls the position of a cursor used to examine the spectral frame being displayed. In this picture, the cursor is positioned at a spectral peak which is

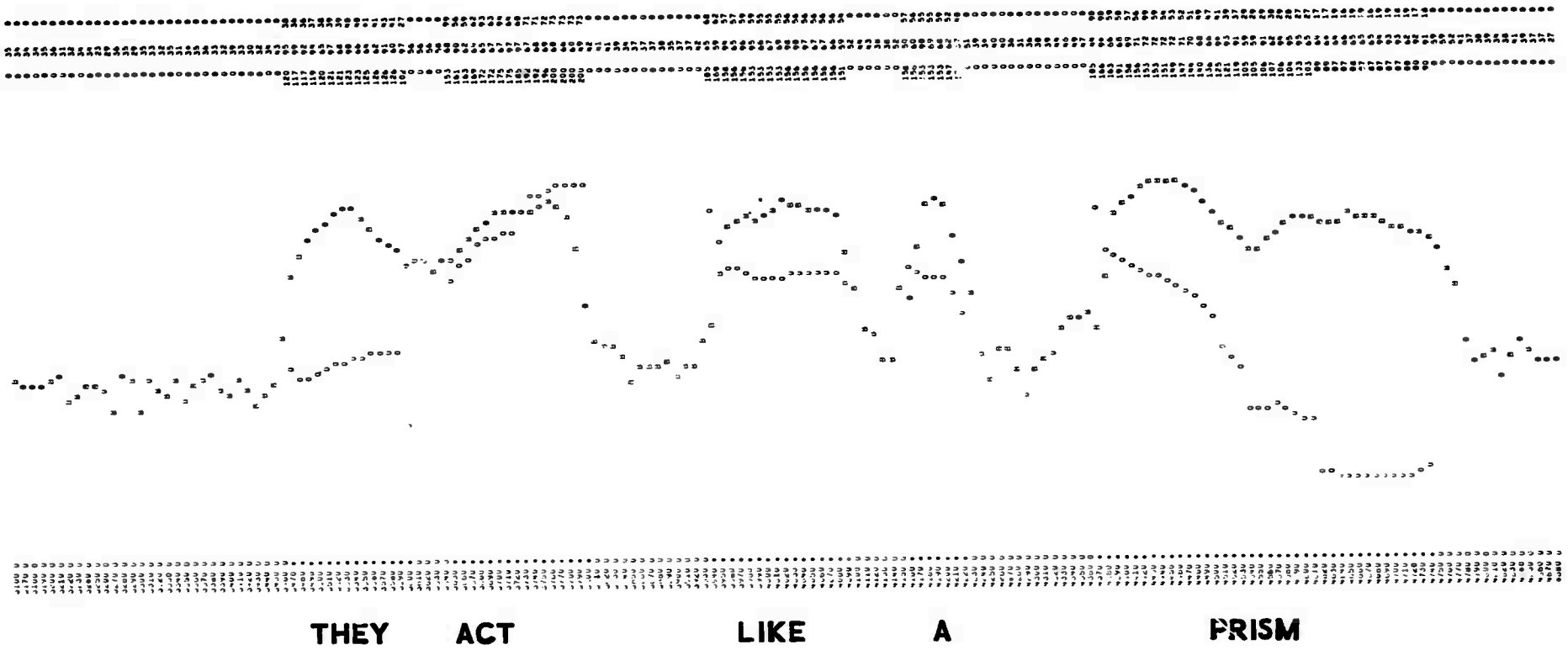



Figure 3. A graph of broadband energy (B) and fundamental frequency (O) versus time for "they act like a prism" by talker ASH. Time goes from left to right in milliseconds, as shown by the set of numbers at the bottom. Immediately above the graph is a tabular listing of fundamental frequency in Hertz, then energy in dB, and at the top, fundamental frequency in eighth-tones (refer to the discussion in section 3.3).

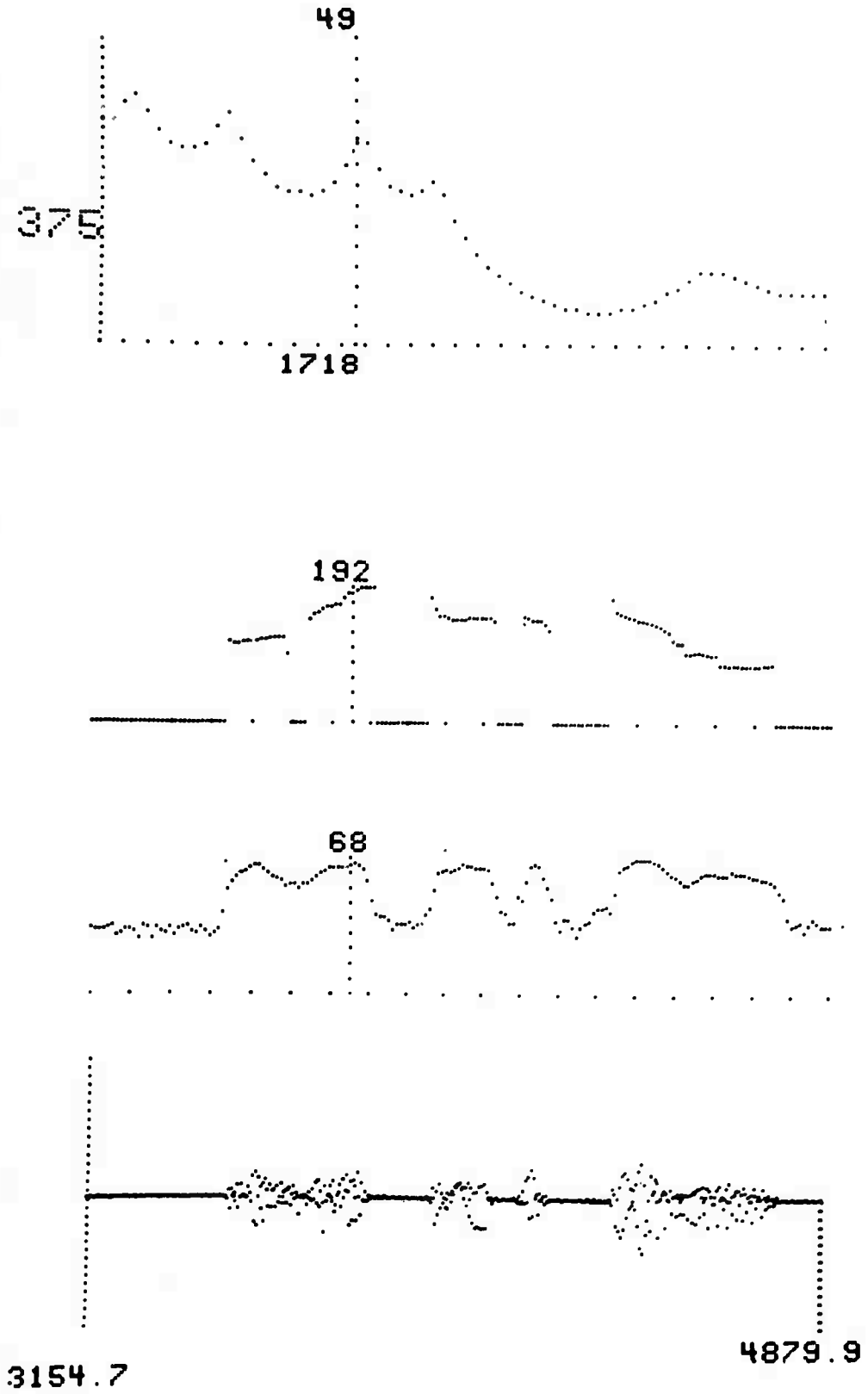



Figure 4. Photograph of the interactive graphical display for the utterance "they act like a prism" by talker ASH. For this example, the display shows (from the bottom up) the time wave, energy contour, fundamental frequency contour and one spectral cross section (refer to the discussion in section 3.3).

centered at 1718 Hz and has a relative amplitude of 49 dB. The middle of the display contains the time functions of fundamental frequency in Hertz and broadband energy in dB. The position of the time function cursors is selected with a potentiometer and, for time 3750 msec, show values of 192 Hz for fundamental frequency and 68 dB for broadband energy.

Figures 3 and 4 thus illustrate the types of displays that can be obtained for speech studies such as the experiments to be described in section 4.

### 3.4 Syntactic Boundary Detection

The prosodic patterns obtained from the interactive research facility can also be used as inputs to programs for detecting aspects of syntactic structure. Recent research (Lea, 1971, 1972a, b) has demonstrated that recognition of aspects of syntactic structure can be accomplished, in part, by using fundamental frequency ($F_0$) contours to detect boundaries between major syntactic units. While for decades linguists have claimed that intonation and stress may indicate the immediate constituent structure of English (Gleason, 1961; Lieberman, 1967a,b; Trager and Smith, 1951; Wells, 1947), this recent work has explicitly demonstrated success in computer detection of syntactic structure.

Fundamental frequency contours were obtained for over 500 seconds of speech, including short stories, newscasts, weather reports, and excerpts from conversations, spoken by nine talkers. A decrease (of about 7% or more) in $F_0$ usually occurred at the end of each major syntactic constituent, and an increase (of about 7% or more) in $F_0$ occurred near the beginning of the following constituent.

Figure 5 illustrates the $F_0$ contour of a typical sentence taken from a weather report. Fall-rise "valleys" (marked by vertical dotted lines) accompany the syntactically predicted boundaries (marked by arrows labelled with the categories of surrounding constituents). A computer program, based on the regular occurrence of $F_0$ valleys at constituent boundaries, correctly detected over 80% of all syntactically predicted boundaries.

Of the less than 20% of "missing" boundaries, about half were due to predicted boundaries between noun phrases and following verbals (auxiliary verbs

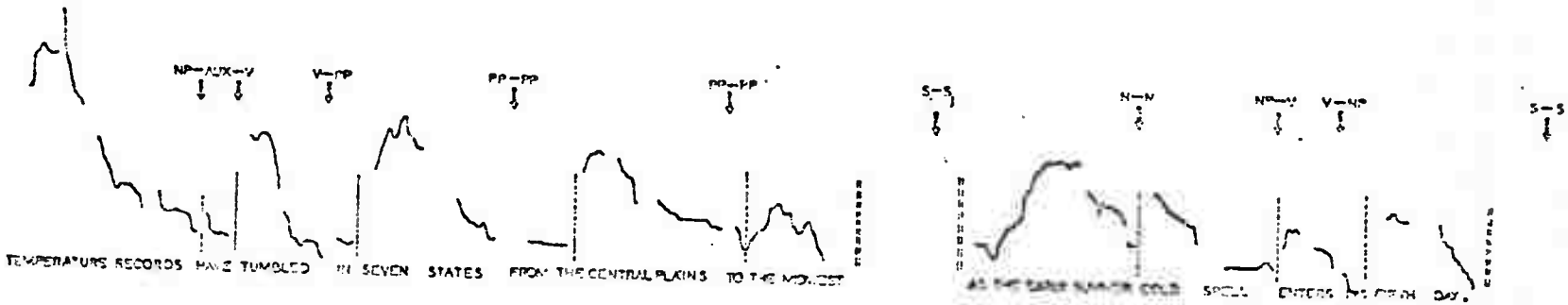





Figure 5. An $F_0$ contour (vertical axis, frequency; horizontal axis, time) of a sentence, with predicted constituent boundaries shown by arrows (labelled with category symbols for surrounding constituents), and detected boundaries shown by vertical lines.

or main verbs). There is considerable evidence (e.g., contractions like "I've, etc.) that such noun phrase - verbal boundaries would not be expected in phonological structure (Lea, 1972a). Thus, when boundaries between noun phrases and verbals are neglected, about 90% of all other boundaries are detected.

Besides this regular acoustic manifestation of boundaries between major syntactic constituents, some boundaries between minor constituents (e.g. between an adjective and a following noun) were also detected by the fall-rise patterns in $F_o$.

Detecting such syntactic structure from $F_o$ contours is complicated by the fact that, at consonant-vowel (and vowel-consonant) boundaries, variations in $F_o$ occur which may be confused with the changes marking syntactic boundaries. False (syntactically unrelated) boundary detections resulted from $F_o$ variations at these boundaries between vowels and consonants, but most such false alarms could be eliminated by setting a minimum percent variation (about 10%) in $F_o$ for a boundary detection. A detailed study of $F_o$ variations at phonetic boundaries (Lea, 1972a, Ch. 4) clearly indicated that such phonetically-dictated changes in $F_o$ would rarely exceed about 10%. Studies were also conducted on the effects of stress patterns on $F_o$ variations at consonant-vowel boundaries.

Sentence boundaries (such as that marked $S_i$ - $S_j$ in Figure 5) were always accompanied by fall-rise $F_o$ contours. In fact, the rise in $F_o$ (around 90% change) after a sentence boundary was substantially larger than the usual rises (about 40% or less) after non-sentential constituent boundaries. In addition, sentence boundaries were usually (in over 90% of all cases) accompanied by long (35 centisecond) stretches of unvoicing. Here "sentence boundaries" refer to both boundaries between matrix (unembedded) sentences and boundaries between embedded full-clausal sentences (as in Figure 5).

A preliminary study was conducted of the effects of specific constituent categories (noun phrase, verb, prepositional phrase, etc.) on boundary detection. The lack of regular boundary marking between noun phrases and following verbals has already been noted. On the other hand, around 95% of all boundaries before

prepositional phrases are detected by $F_o$ fall-rise valleys. This might be especially useful, since NP-PP-PP sequences are known to give particular difficulties to syntactic parsers. Also, coordinate noun phrases or coordinate adjectives were always accompanied by $F_o$ valleys between the conjuncts. Sizes of $F_o$ valleys were also studied for the various syntactic categories.

The constituent boundary detection program developed by Lea at Purdue has been implemented as a FORTRAN program at the Univac DSD Speech Communications Laboratory. The experiments with $F_o$ - detected constituent boundaries will be extended to other texts and talkers (see section 4). To further refine the studies of how syntactic category affects $F_o$ contours, a controlled experiment is also being planned, in which position in the sentence, constituent category, lexical content, and other factors can be varied separately in designed texts. Syntactic contrasts, such as compound structures versus nuclear structures, or simple constituents (NP: John) versus coordinate structures (NP and NP: John and Bill), etc., would be placed at various points in the structure of a sentence to isolate the effects that syntactic categories and bracketing may have.

The previous studies of boundary detection have been confined to declarative sentences in spoken texts and to declarative and (a few) imperative utterances excerpted from interviews. Since man-machine interactions for information retrieval, or for other tasks discussed by ARPA Speech Understanding Research contractors, will undoubtedly involve commands and questions, investigations of boundary detection in questions and commands would be appropriate. This may introduce the need for refined boundary detection techniques to handle other types of sentences.

Such studies also can be extended by investigating what syntactic information can be extracted from speech intensity contours and phonetic durations (Willems, 1972). In particular, intensity sometimes drops at boundaries, much as $F_o$ does (though sometimes more dramatically). Intensity contours can also give more precise specifications of silent pauses than the mere absence of $F_o$ can. Phonetic durations are lengthened in phrase-final positions (Allen, 1968; Barik, 1969; Barnwell, 1970; Boomer, 1965; Goldman-Eisler, 1958; 1961), but the use of durations requires phonetic segmentation processes.

To date, studies have not been concerned with exact boundary location; only boundary detection. When weakly stressed or reduced syllables begin a constituent, the $F_o$ valley bottom may occur within that weak beginning of the constituent. When a previous constituent exhibits a "Tune II" intonation rise (Armstrong and Ward, 1926) at its end, the $F_o$ valley bottom may occur within the last syllables of the prior constituent. Refinements might be incorporated to more exactly locate the boundary within the region of the $F_o$ valley.

The refined procedures for syntactic segmentation must be integrated with stressed-syllable location procedures. To develop clear indications of the acoustic correlates of stress in connected speech, and to test refined segmentation procedures, the experiments described in section 4 have been devised.

## 4. EXPERIMENTS ON PROSODIC PATTERNS IN THE RAINBOW SCRIPT

The philosophy of speech recognition outlined in section 2.4 suggests the need for methods of demarcating constituents, finding stressed syllables, and doing a partial distinctive features analysis on the reliable data within the stressed syllables. A method for demarcating constituents was outlined in section 3.4. Its implementation and refinement at Univac must be tested with extensive speech data. Methods for finding stressed syllables, and for refining partial distinctive feature estimation techniques will be developed. They depend upon first finding reliable acoustic correlates of stressed syllables.

Stress is an abstract quantity usually considered to be associated with a speaker's total physical effort in speech production or with a listener's perception of "prominent" syllables. Extensive work has been done on acoustic correlates of stress (cf. reviews by Lehiste, 1970, and Medress and Skinner, 1972), and on physiological correlates of stressed syllable production (cf. review by Lieberman, 1967). On another hand, linguists have devised phonological rules that purport to predict the stress levels and vowel reductions in spoken English (Chomsky and Halle, 1968; Halle and Keyser, 1971). Yet, this work has not answered vital questions about how to automatically locate stressed syllables in connected speech.

In this section, we will outline experiments which are designed to determine what acoustic features correlate well with the stress levels and syntactic boundaries in a connected speech text. Based on the experimental results to be obtained, computer-implementable techniques for stressed syllable location (and refined constituent boundary detection) will later be developed.

A three-fold experimental effort is involved in this research, with these major data-gathering tasks:

1. Syntactic analysis of the sentences in the speech text, followed by application of appropriate stress rules and vowel reduction rules, to predict stress levels and vowel reductions in the script;
2. Presentation of the script, spoken by six talkers (4 male, 2 female), to four listeners (individually), for their judgments as to which syllables are stressed, reduced, or unstressed; and

3. Processing of the spoken scripts by the interactive speech research facility (see section 3), to obtain data on the $F_0$ contours, intensity contours, and other acoustic features that may correlate with syllable stress and reduction.

Following such data-gathering tasks, there come the extensive tasks of: relating the linguistic predictions of stress to the perceptual results; comparing the perceptions of the various listeners, and determining variations from talker to talker; and relating the various acoustic features to the perceptually established stress patterns and to the linguistically predicted stress patterns. Conclusions must then be drawn concerning; the adequacy of the linguistic rules in predicting listener judgments; the regularity of stress judgments, from talker to talker and from listener to listener; and the best acoustic correlates of stress and reduction. Methods for automatically predicting stress from acoustic correlates must be considered.

In section 4.1 the overall design of the experiments is discussed and related to previous studies of stress. The methods of syntactic analysis, and the predictions of stress that will result from linguistic stress rules, are discussed in section 4.2. Section 4.3 describes the study of listeners' perceptions of stress patterns in the spoken text. Acoustic correlates of perceived or predicted stress are discussed in section 4.5.

## 4.1 Selection of Experimental Conditions

The text chosen for the initial studies of stress analysis and boundary detection is the first paragraph of the "Rainbow Passage" introduced by Grant Fairbanks (1940), and used extensively in speech research. The text, hereinafter referred to as the "Rainbow Script", reads as follows:

> "When the sunlight strikes raindrops in the air, they act like a prism and form a rainbow. The rainbow is a divisions of white light into many beautiful colors. These take the shape of a long round arch, with its path high above, and its two ends apparently beyond the horizon. There is, according to legend, a boiling pot of gold at one end. People look, but no one ever finds it. When a man looks for something beyond his reach, his friends say he is looking for the pot of gold at the end of the rainbow."

Several criticisms of previous work motivate the present experiments with the Rainbow Script. These criticisms will be considered in the following numbered paragraphs. Numbering of the criticisms here corresponds with the similarly numbered advantages of the present experiment, to be tabulated at the end of this section.

(1) Some studies of acoustic correlates of stress (e.g., Medress, Skinner, and Anderson, 1971) have been based on intuitive knowledge of stress patterns, while others (e.g., Fry, 1958, Bolinger, 1958; Morton and Jassem, 1965) have attempted to determine how listeners' judgments of stress vary as certain acoustic features are varied. These studies of performance may be contrasted with linguist's theoretical models (i.e., sets of stress-assignment rules), which are usually based on the competence of an ideal speaker-listener, distinguished from performance by the exclusion of memory-limitations, speaker and listener differences, etc. The present study attempts to associate linguistic predictions, listener judgments, and acoustic features. From such a three-way association we may learn how acoustic features correlate with both ideal and actual listener-assigned stresses, we may see something about speaker and listener variability from a theoretical norm, etc.

(2) Many studies of acoustic correlates of stress have been based on presenting synthesized speech to listeners. Such studies take advantage of the ability to separately control acoustic features of synthesized speech, for testing how acoustic variations correlate with listeners' perceptions of stress (Fry, 1955, 1958; Bolinger, 1958; Morton and Jassem, 1965; Mattingly, 1966; etc.). However, one must be very cautious about simply extrapolating from results with unnatural synthetic speech to corresponding claims about natural speech.

Similarly, studies of listeners' perceptions of structural boundaries and other prosodic structure have been attempted with speech data distorted so as to remove or mask all or most of the segmental phonetic information, leaving only prosodic information for the listener (O'Malley and Peterson, 1966; Blesser, 1969; O'Malley, 1972; cf. also Lummis, 1971). Techniques used include inverting the frequency spectrum, masking with noise, and low-pass filtering. While such distortions may, with some difficulty, substantially isolate prosodic information

from segmental data and lexical context information, the listeners' behavior with such distorted speech may or may not correspond well with their responses to prosodic patterns in natural speech. Studies with natural speech would ultimately seem appropriate.

(3) A recurrent problem in stress studies is the indiscriminate confusion of stress and emphasis, and the loose concept of exactly what stress is. The most blatent violations of "knowing what you're looking for before you seek its acoustic correlates" occur when researchers ask for listener's judgments about stress while they take a fixed unambiguous utterance and increase $F_0$, intensity, vowel durations, or such within one vowel or syllable. Thus, for example, Bolinger (1958) worked with individual utterances such as:

Wouldn't it be easier to wait?
Break both apart.
Many are taught to breathe through the nose.
But would many return?
Alexander's an intelligent conversationalist.

and varied acoustic features, asking listeners to judge whether easier or wait was more stressed, etc. Lieberman (1967, Ch. 4) studied the contrast between such sentences as

Joe ate his scup.
Joe ate his soup.
Joe ate his soup.

where the underlined word is given special prominence or emphasis. Such studies deal with what properly may be called special emphasis, in that words or syllables are assigned a prominence or force of utterance which is non-normative, marking a semantic distinction from other sentences which do not exhibit these special effects. In such utterances, prominence is specifically intended to mark a distinction from the norm, or semantically neutral utterance, with the same word content.

In contrast to such emphatic prominence, stress is used in this report to refer to the relative prominence of syllables in the normative utterance of a sentence. This normative stress pattern, which we might also term linguistic stress, is what should be predicted by linguistic stress assignment rules, such as those of Halle and Keyser (1971). The performance of an individual speaker, or the judgments of an individual listener, will approximate this norm, but will be influenced by extralinguistic factors.

Many studies (Fry, 1955, 1958; Bolinger, 1958; Lieberman, 1960; Mol and Uhlenbeck, 1956; Morton and Jassem, 1965; Lea, 1972, Ch. 5) have investigated the acoustic correlates of stress contrasts in isolated minimal pairs such as noun-verb pairs (permit - permit, etc.). The semantic and syntactic distinctions in such pairs are marked by stress contrasts, not phonemic sequence differences. While such studies help isolate stress effects from other acoustic factors, they involve a special case of stress effects which may or may not give acoustic cues which correspond with those given in connected sentences or even in other multi-syllabic non-minimum-pair words or phrases.

(4) While linguists (Trager and Smith, 1951; Chomsky and Halle, 1968) have often worked with four or more levels of stress, plus a category of reduced vowels or syllables, some tests show that trained listeners can reliably and consistantly judge no more than two stress levels, plus identifying reduced syllables (Lieberman, 1964; Lea and Li, IN PREPARATION). Three levels of stressedness will be used in the present perceptual or acoustic studies: stressed, unstressed, and reduced.*

(5) The present study is concerned with sentence stress, not word stress. While word stress (also called "lexical stress") is one form of input information into the rules for assigning sentence stress, the overall sentence stress pattern is also a function of syntactic bracketing and syntactic categories. Few experimental studies have been concerned with the stress patterns throughout sentences. Previous perceptual tests with sentence material have involved deciding which is the most stressed syllable, whether a specific single syllable is or is not stressed, or which of two syllables is more stressed. The present experiments extend studies to all syllables in the sentences.

*These stress categories are defined in section 4.3. An exception, noted there, is where one listener (WAL) marked four levels of perceived stress for one repetition of the perceptual experiment.

(6) (7) (8) Isolated sentences tend to have different intonation contours, and perhaps different stress patterns, from those in sentences in the context of discourse. The Rainbow Script used in the present experiments is a well-known semantically connected text (a paragraph) with substantially neutral content, demanding few (if any) cases of special emphasis, but with a variety of declarative structures included. Compound nouns, nuclear phrase structures, parenthetical-like phrases, conjuncts, and complex sentence embeddings are all exhibited. Interrogatives which exhibit rising $F_o$ in sentence-final syllables are avoided. This is one confusion factor which questions like Bolinger's (illustrated above) introduced into previous studies. Declaratives, imperatives, yes-no questions, and WH-word questions should be distinguished and handled separately in stress studies.

(10) Intonation of various sentence types has the most pronounced effect on the last stressed syllable of a sentence (or clause) and any subsequent unstressed syllables (cf. Lehto, 1971; Armstrong and Ward, 1929). Figure 3 in section 3.3 (p. 25), illustrates several effects of clause-final tonalization positions. In the clause "they act like a prism" of Figure 3, act and 'pris' of prism are both stressed. However the clause-final word prism is much longer, and shows a characteristic falling $F_o$ contour, in contrast to a shorter and rising-$F_o$ syllable act. High peak $F_o$ would suggest that act was stressed, while long duration of prism may either be attributed to stress or the clause-final position (which lengthens both stressed and unstressed syllables; Mattingly, 1966). Consequently, perceptual and acoustic results in these clause-final positions (called tonalization positions) may be different from those in the rest of the sentence. The analysis of Rainbow Script data should involve a separate consideration of acoustic features of stress in the tonalization and non-final (so-called intonation "body") positions.

(11) (12) Another significant aspect of the present study is the set of acoustic features to be monitored. Several different parameters of $F_o$ contours, both within vowels and within consonants, are to be compared with predicted and perceived stress patterns. Similar parameters of energy contours are also to be studied, for both broadband energy and low-frequency energy. Durational measures, while recognized to be closely correlated with stress, are deemphasized in this study. The reason is that the required phonetic boundaries

and vowel and consonant durations are difficult to automatically determine. The potential for machine extraction of such acoustic cues is given special consideration here, since the intention is to develop stress cues suitable for application in speech recognition systems.

(13) In evaluating acoustic correlates of stress, the influences of phonetic environment (vowel features and pre- and post-vocalic consonants) on $F_o$ contours, energy contours, and durations must be considered. Recent research (Lea, 1972a, Ch. 5) showed that, in isolated words, a falling $F_o$ contour in the beginning of a vowel may indicate either that the preceding consonant was unvoiced or that the syllable was unstressed. A rising contour may indicate a preceding voiced consonant, a word-initial vowel, or stress on the syllable. Peak $F_o$ and amplitude in a vowel are affected by whether the surrounding consonants are voiced or unvoiced, and by whether the tongue is high or low in the oral cavity during the vowel (Lehiste, 1970; Lea, 1972, Ch. 4). Vowel durations are lengthened before voiced consonants (House and Fairbanks, 1951; House, 1960; Lehiste, 1970). The study of acoustic correlates in the Rainbow Script will include consideration of such phonetic effects.

(14) The perception tests to be reported in section 4.3 are with several listeners and several talkers chosen to be representative of a wider population (essentially, those with General American dialect). One listener repeated the perception test several times to determine how consistent stress perceptions are from time to time.

(9) (15) Some studies of listeners' perceptions of stress (e.g., Bolinger, 1958) have required the listener to simply record whether or not a given syllable is stressed, or, alternatively, which of two syllables is more stressed than another. To get judgments for all the syllables in a sentence, the task would have to be repeated, once again with each other syllable as the one in question. An alternative procedure is to have the listener listen repeatedly to the same recorded speech until he was able to assign a judgment to each syllable. This is the technique used in the perceptual tests of the present study. The method of repeatedly listening to each utterance has apparently

not been used before in stress perception studies, and its relative merits are not established. Two similar techniques were employed by listeners in this study; one whereby the tape is rewound and replayed at will, and another where a sentence or clause of speech is digitized and repeatedly replayed by computer until the listener terminates the repetition.

In summary, we may list the following characteristics of the present experiments with the Rainbow Script:

(1) The experimental design incorporates linguistic theoretical conclusions about stress patterns, listeners' perceptions of stress, and studies of acoustic correlates, all in one effort.

(2) The speech is natural, not synthetic.

(3) The study is of linguistic stress (and reduction), not special emphasis or special minimal-pair differences (as with noun-verb pairs).

(4) Three levels of stressedness are studied: stressed, unstressed, and reduced.

(5) Sentence stress, not word stress, is being studied.

(6) A semantically-connected text (paragraph) is used, rather than isolated sentences.

(7) The text is a well-known text, used extensively in speech studies, and originally designed to display "habitual pitch patterns" (Fairbanks, 1940).

(8) The text incorporates one type of sentence (declarative), with a variety of phrasal structures.

(9) Perception judgments are provided for all syllables; no forced choices are demanded as to "most stressed syllable in utterance" or "syllable A is more stressed than syllable B".*

(10) In the analysis of perceptual and acoustic results, sentence tonalization (clause-final or breath-group-final) positions are analyzed separately from neutral (non-final; intonation body) positions.

(11) Extensive sets of acoustic parameters are considered as prospective acoustic correlates of stress.

---

*In measurement-theory terms, the judgments here form a nominal scale, rather than an order or ordinal scale of measurement (Stevens, 1951; Lea, 1971).

(12) In evaluating acoustic correlates of stress, consideration is given to both the effectiveness in correlating with (or "predicting") stressedness or reduction and the potential for automatically extracting the acoustic features from connected speech.

(13) The influences of phonetic environment (vowel features, voiced and unvoiced post- and pre-vocalic consonants) are considered in evaluating acoustic correlates.

(14) Perception tests on stress are made by several listeners, hearing the utterance portions repeatedly, with several male and female talkers, and with a test made of the repeatability of listener judgments.

(15) Tape rewind and replay, as well as computer storage and replay, provide two distinct methods of speech presentation to the listener, which may be compared.

(16) Latest work on English stress rules and vowel reduction rules is applied to the performance problems of predicting listener perceptions of stress and reduction, and of acoustic correlates of linguistic predictions.

In brief, this set of experiments provides a comparison among linguistic, perceptual, and acoustic data on total stress patterns in a well known connected declarative discourse, spoken by several talkers. Special attention is given to interfering effects of intonation, phonetic context, speaker and listener differences, etc.

In section 4.2, we elaborate on advantage (16) in the above list, pointing out how recent linguistic work will be applied to practical stress predictions.

## 4.2 Syntactic Analysis and Linguistic Stress Predictions

Recent published rules suggest that stress patterns within words and in total sentences can be predicted by stress rules (and vowel reduction rules) which require phonemic content, syntactic bracketing, and syntactic category names as input information. To predict stress patterns using such rules, one must first perform a syntactic analysis of the script, and specify those phonemic aspects that are relevant to lexical stress assignment.

A complete syntactic analysis of the Rainbow Script will be done, using a transformational grammar. Deep syntactic structure will be found for each sentence in the script, and the surface structure will be determined by applying that deep structure to an ordered set of transformations (based on

an extension of the grammar given in Burt, 1971). The resulting surface structure will then to be subjected to phonological readjustment rules (Chomsky and Halle, 1968, pp. 24f) to yield the so-called "phonological representation" of the sentences. No adequate set of phonological readjustment rules have been written, but some expected effects are known. A decision must be made either to hypothesize what phonological representation will result from the rules, or to design adequate rules. The phonological representation is used to predict expected constituent boundaries, for comparison with boundary detector results (cf. Lea, 1972a, Appendix B.)

The phonological representation will also include phonemic-string information for the words in the structures. This phonological representation is then to be processed through stress assignment rules and vowel reduction rules, such as provided by Halle and Keyser (1971; based on revisions of rules given in Chomsky and Halle, 1968; cf. also Ross, 1969, and Vanderslice and Ladefoged 1971). Selection of adequate stress and reduction rules is another major task. Recent revisions in stress assignment based on putting the Nuclear Stress Rule within the syntactic transformational cycle must also be considered (Bresnan, 1971, 1972; Lakoff, 1972; Berman and Szamosi, 1972). Vowel reduction rules must be given similar attention, and predictions of reduced vowels must be obtained as well as predictions of stress levels. Alternative rules for assigning stress and reduction can be compared with the perceptual results, to determine which rules seem to be most satisfactory (cf. Trager and Smith, 1951; Crystal, 1969).

At the time of this writing, study of appropriate syntactic and phonological rules has just begun, and linguistic predictions are thus yet to be obtained.

### 4.3 Perception Tests

The Rainbow Script has been read by six talkers, providing natural speech for both perception and acoustic analyses.* Four listeners have been presented with the speech, and asked to record, for each syllable, their individual judgment as to whether the syllable was spoken as stressed, unstressed, or reduced. One listener repeated the experiment several times to establish listener consistency

*We are indebted to George W. Hughes and Kung-Pu Li at Purdue University for providing the speech recordings and some of the perceptual data for the Rainbow script. The perceptual testing procedures used here are based on a modification of the procedures used by Hughes, Li, and Snow (1972).

and effects of some other experimental variables. Here we discuss in some detail the preliminary results of these perceptual tests.

The basic method of the perceptual study consists of presenting, to an individual listener through earphones, the recordings of each of the six talkers. The listener is asked to mark (in whatever way he chooses), for each syllable, whether he hears that syllable as stressed, unstressed, or reduced. To facilitate marking for each syllable, the script is typed on a sheet of paper with vertical slashes between syllables. A mark must be provided for each syllable (between two slash marks). The listener receives one such sheet for each talker. Results were obtained for listeners (ASH, GWH, WAL, and TBS) who are trained in the speech field, and thus in some sense familar with notions of stress and vowel reduction.

The script was spliced into clause or sentence portions separated by long (about 3-second) pauses. This facilitated stopping the recording after each clause, recording certain judgments, then backing up the recording again to the beginning of that portion of the script, to hear to again. The listener could listen to the tape portions as often as necessary to mark each syllable. He was free to back up the tape at his choice, and no time limit or procedural constraints were placed on him. At least one listener (WAL) found that approximately once for each syllable a tape rewind and listening was required to firmly establish the categorizations.

Listeners reported that some syllables were clearly stressed and some clearly reduced, while many were not so readily categorized. In an initial experiment where listeners had been asked to mark each stressed syllable and each reduced syllable, those not marked were, by default, considered as unreduced, unstressed. This bias toward the extremes of high stress and lowest stress (reduction) probably carried over into the final experiment. Two binary decisions thus may appear to be involved in the judgments: "Is the syllable stressed?" and "Is the syllable reduced?" A "no" answer to each yields the middle ground of "unstressed" syllables.

Figure 6 illustrates the results of the initial perception test for one talker (ASH). Perceptual results from a fifth listener (RPS) were included along with those of the four listeners previously mentioned. Plotted for each of the syllables in the Rainbow Script are the number of stressed judgments minus the number of reduced judgments, for the five listeners. Unstressed judgments were assigned values of zero. (No cases occurred where one listener's reduced judgment cancelled another's stressed judgment.) Thus, if all five listeners heard a syllable as stressed, a value of 5 was plotted; if two perceived a syllable as reduced, and the other three perceived it as unstressed, a value of -2 (minus two) resulted. The syllables which were most definitely stressed (i.e., perceived by all listeners as stressed) thus were at the top of the scale; those definitely perceived reduced were at the bottom of the scale. One listener (ASH) unfortunately provided no judgments about occurrences of reduced syllables. Thus, the most negative values shown are -4, indicating unanimous agreement among the four listeners judging reduced syllables.

Figure 6 thus shows which syllables are judged by a set of listeners to be more or less definitely stressed, unstressed, or reduced. While these results are for the initial test where every syllable did not have to be marked (so that unmarked syllables were, by default, considered unstressed), similar results are to be obtained for the more controlled tests where each syllable is categorized. From such results, one can readily see which syllables are unanimously judged as stressed, which are judged as stressed by a majority of the listeners, etc. When syllables such as long, round, arch in the second sentence shown in Figure 6 are unanimously judged as stressed, one can be more confident that acoustic cues to stress are to be found.

The results of one listener (WAL) marking his perceptions for the same talker (ASH), under several different conditions, are shown in Figure 7. As in Figure 6, the judgments are "plotted" for each syllable. The little boxes connected by dashed lines show whether the listener judged the syllable as stressed, unstressed, or reduced, when the listener was required to mark all syllables, while repetitively rewinding and replaying the taped speech. Also shown in Figure 7 are X's marked at the stress level judged in the earlier test where no mark was provided for unstressed or missed syllables. These X's are included only

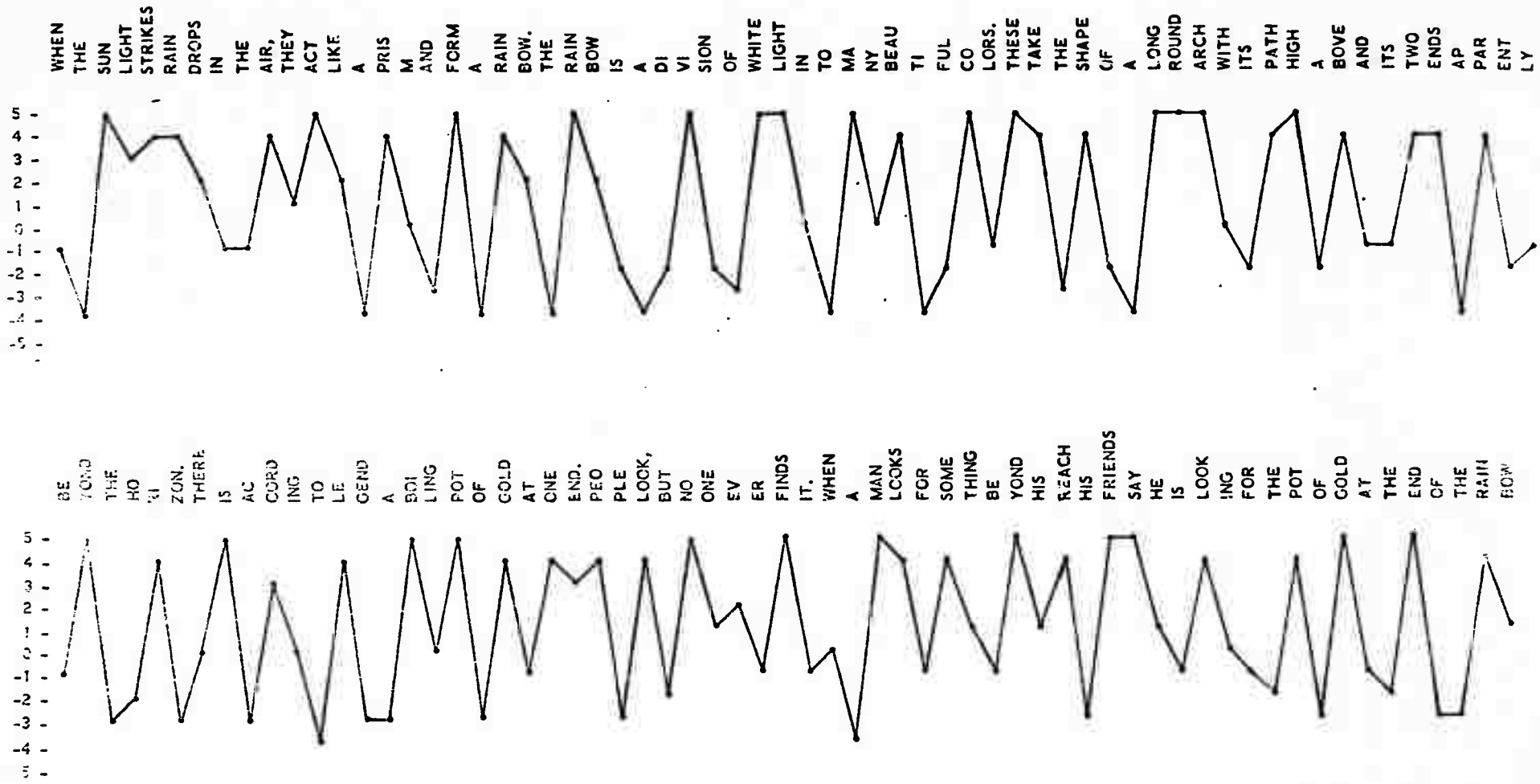



Figure 6. Summary of stress judgments by five listeners, for one talker (ASH) reading the Rainbow Script. Plotted for each syllable is the number of judgments of the syllable as stressed minus the number of judgments of the syllable as reduced. Unanimous judgment as stressed thus yields the top value of +5, whereas judgments as reduced pull the value down toward -5.

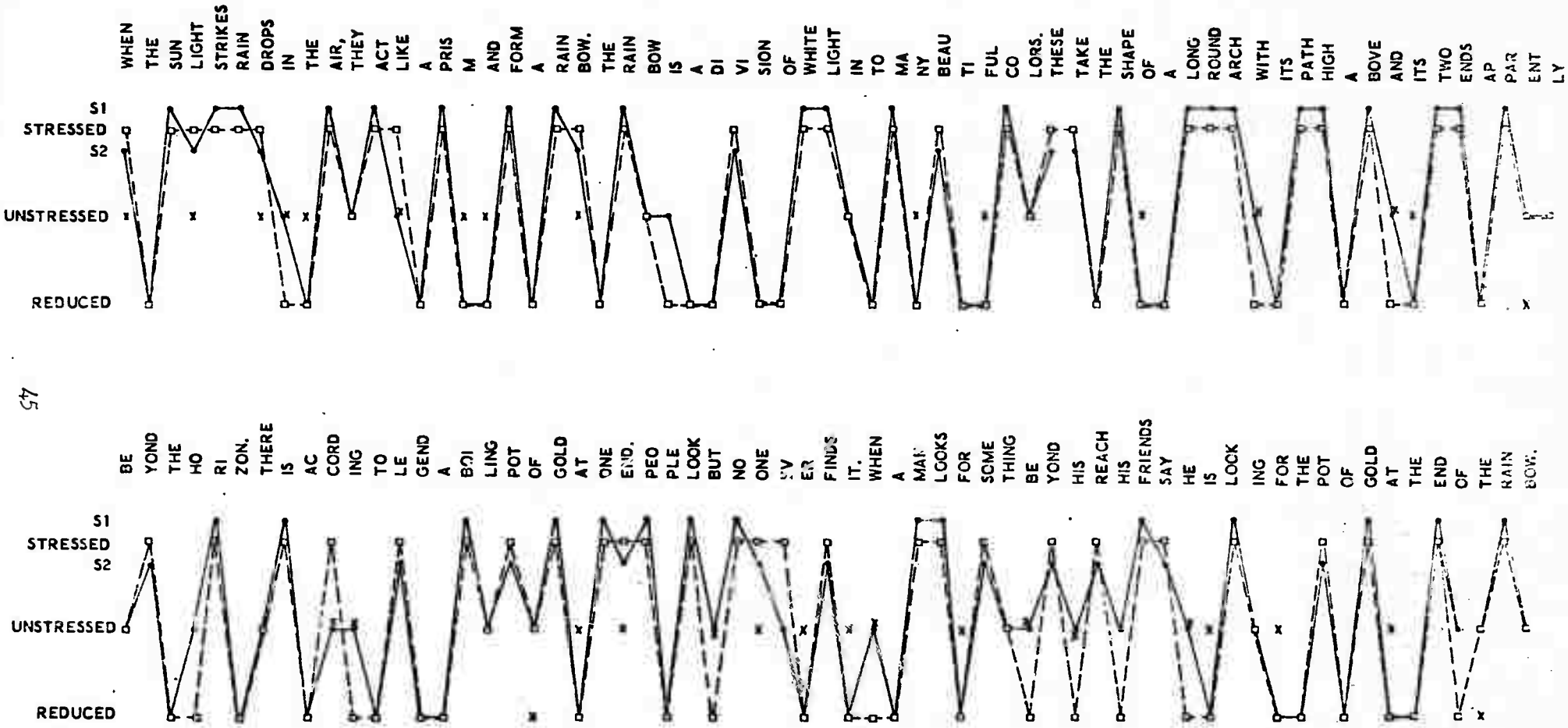



Figure 7. Stress judgments from several repetitions of the perception test by one listener (WAL), with the Rainbow Passage read by one talker (ASH). Solid lines connect judgments on a four-level scale of highly stressed ($S_1$), lesser stressed ($S_2$), unstressed, or reduced, obtained when the computer repetitively replayed the clauses of the speech. Dotted lines connecting between the boxes mark three-level results (stressed, unstressed, or reduced) for a tape-rewind-and-replay approach. Wherever judgments on another run of the three-level judgment differed from the boxed values, x's are shown. This third run was conducted with the listener marking only stressed and reduced syllables (see section 4.3).

where the earlier results differed from the judgments when every syllable had to be marked. The preponderance of X's at the "unstressed" level shows that most of the differences from the later, complete test were due to missing marks on what may well have been perceptually stressed or reduced syllables. This supports the need for requiring expressed judgments on all syllables.

Figure 7 also shows results when the same listener used the Univac interactive speech research facility to digitize and store the clausal portion from the tape, and play it back repeatedly through the D/A converter. This eliminated the cumbersome stopping, rewinding, listening, stopping, rewinding, etc. of the tape recorder system. (It did, however, introduce some background noise due to A/D and D/A processes.) The ready repetition with the digital system allowed a finer judgment of stress "levels", the listener believing he could then separate stressed syllables into two categories: highly stressed and less stressed. These four categories are shown in Figure 7, and the judgments for all syllables are connected by the solid lines. Comparing all these results syllable-by-syllable, it is evident that the two tests give similar results, with almost all "highly stressed" syllables and most "lesser stressed" syllables from the computer-aided test corresponding to stressed syllables in the tape-replay test. The four-level judgments might be said to break up a continuum (from highest-stressed to reduced) into four categories, by a different setting of 'thresholds' than the three-level judgment involves.

Figure 7 is drawn to such a scale that it could be laid directly over Figure 6, to show a close correspondence between the results of five listeners reporting on one talker and the results of one listener reporting on the same talker under several conditions. The close correspondence between the syllable-by-syllable judgments plotted in Figures 6 and 7 shows that the one listener is in one sense "representative" of the group of listeners.

Figure 8 summarizes the perceptions of listener WAL when marking all syllables in the scripts read by each of the six talkers. Results with the tape-rewind approach are compared with those for the four-level results of the computer-replay approach. As evidenced by the syllable-by-syllable 'plot' of Figure 7 (for one talker), those syllables judged as highly stressed in the computer-replay

| TRIAL W/COMPUTER \ PERCEIVED STRESS LEVEL → | TRIAL WITH TAPE REWIND: STRESSED | UNSTRESSED | REDUCED |
|---|---|---|---|
| HIGHLY STRESSED | 36 | 0 | 0 |
| LESSER STRESSED | 21 | 0 | 0 |
| UNSTRESSED | 3 | 14 | 14 |
| REDUCED | 0 | 0 | 39 |

(a) Talker ASH

| TRIAL W/COMPUTER \ PERCEIVED STRESS LEVEL → | TRIAL WITH TAPE REWIND: STRESSED | UNSTRESSED | REDUCED |
|---|---|---|---|
| HIGHLY STRESSED | 34 | 0 | 0 |
| LESSER STRESSED | 19 | 2 | 0 |
| UNSTRESSED | 4 | 14 | 10 |
| REDUCED | 0 | 4 | 40 |

(b) Talker GWH

| TRIAL W/COMPUTER \ PERCEIVED STRESS LEVEL → | TRIAL WITH TAPE REWIND: STRESSED | UNSTRESSED | REDUCED |
|---|---|---|---|
| HIGHLY STRESSED | 32 | 0 | 0 |
| LESSER STRESSED | 23 | 2 | 0 |
| UNSTRESSED | 2 | 15 | 18 |
| REDUCED | 0 | 2 | 33 |

(c) Talker WB

| TRIAL W/COMPUTER \ PERCEIVED STRESS LEVEL → | TRIAL WITH TAPE REWIND: STRESSED | UNSTRESSED | REDUCED |
|---|---|---|---|
| HIGHLY STRESSED | 26 | 0 | 0 |
| LESSER STRESSED | 29 | 2 | 0 |
| UNSTRESSED | 2 | 24 | 12 |
| REDUCED | 0 | 4 | 28 |

(d) Talker JP

| TRIAL W/COMPUTER \ PERCEIVED STRESS LEVEL → | TRIAL WITH TAPE REWIND: STRESSED | UNSTRESSED | REDUCED |
|---|---|---|---|
| HIGHLY STRESSED | 36 | 0 | 0 |
| LESSER STRESSED | 18 | 2 | 0 |
| UNSTRESSED | 1 | 18 | 11 |
| REDUCED | 0 | 4 | 37 |

(e) Talker PB

| TRIAL W/COMPUTER \ PERCEIVED STRESS LEVEL → | TRIAL WITH TAPE REWIND: STRESSED | UNSTRESSED | REDUCED |
|---|---|---|---|
| HIGHLY STRESSED | 37 | 0 | 0 |
| LESSER STRESSED | 20 | 1 | 0 |
| UNSTRESSED | 1 | 19 | 8 |
| REDUCED | 0 | 3 | 38 |

(f) Talker ER

Figure 8. Comparison of three-level and four-level stress judgments of one listener (WAL) marking all syllables in the Rainbow Script spoken by each of six talkers. The three-level test was run with the tape rewind method, while the four-level test involved computer storage and replay. Tabulated in each matrix position is the number of syllables categorized as shown by the headings on the respective row and column.

run (top rows in all matrices of Figure 8) were usually judged as stressed in the tape-replay run (left most column). Those judged as stressed, but at a lesser level, (second row) in the computer-replay run were usually judged as stressed in the tape-rewind run. Most that were judged as reduced in one run were judged as reduced in the other. Thus, results for each of the other five talkers were similar to those reported in detail in Figure 7 for single talker ASH. In this sense, talker ASH is representative of the other talkers.

Figure 8 also shows that, of the 127 syllables in the Rainbow Script, about 50 to 60 (about 40 to 50%) were judged as stressed (or, alternatively, "highly stressed" or "lesser stressed"), slightly fewer (about 35 to 40%) were judged as reduced, and only about 14 to 30 (10 to 25%) were judged as unstressed. Thus, if one were to analyze only the stressed syllables, as suggested in section 2.4, the distinctive-features analysis could be avoided in the 50 to 60% of unstressed and reduced syllables, where distinctive features analysis is most difficult and unreliable. Figure 8, and Figures 6 and 7 as well, also illustrate that more confusions or inconsistencies occur between unstressed and reduced categories than between stressed and unstressed syllables (cf. Lehto, 1969). Thus, a procedure for reliably distinguishing stressed from unstressed syllables might be more successful than one for distinguishing unstressed from reduced syllables.

Figure 9 shows perception comparison matrices for directly comparing listener WAL's judgments for talker ASH with those for the other five talkers. The entries off the main diagonal of each matrix show the deviations from identical perceptions for talker ASH and the other talker. Since a large majority of the syllables were either perceived as stressed for both talkers, unstressed for both, or reduced for both, talker ASH is representative of the other talkers. A study of the perceptions for each of the individual syllables (as Figure 7 provided for one talker and one listener) will indicate which syllables are most likely to be pronounced differently by different talkers, and which syllables have the most stable stress assignment.

Perception tests by the three other listeners are now being conducted, with each syllable to be marked as stressed, unstressed, or reduced. Data such

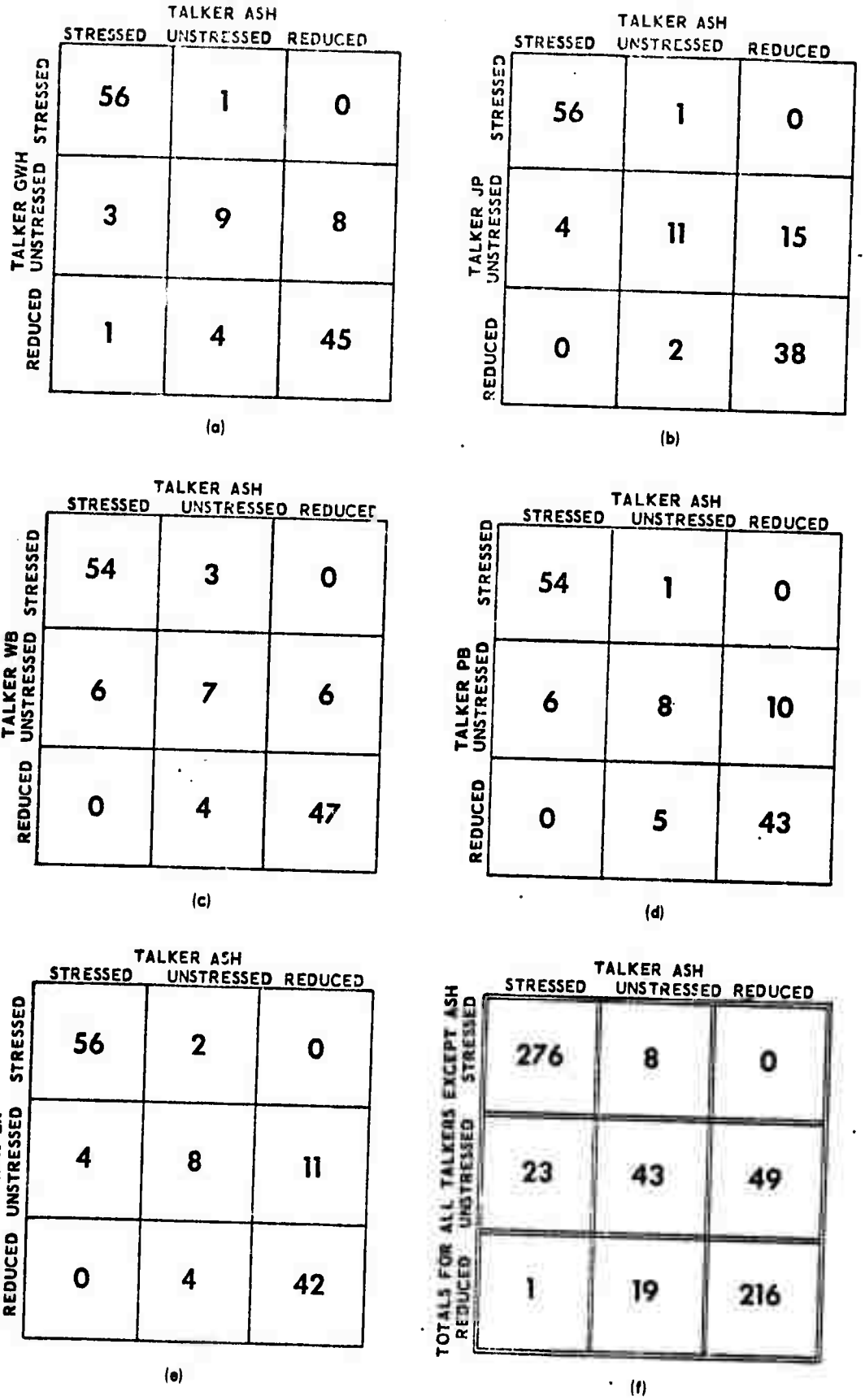



Figure 9. Comparsions of stress level perceptions by listener WAL for talker ASH versus the other five talkers. Tabulated are the number of syllables judged as shown by the column heading, for ASH, and judged as shown by the row heading for the other talker. Figure (f) shows the sum of the results in the five matrices of (a) to (e).

as that illustrated in Figures 6 to 9 will be obtained for these additional tests.

The perceptual results are to be compared with the linguistically predicted stress levels to see what abstract levels are judged as stressed, unstressed, or reduced. For example, are stress levels 1 and 2 both perceived as "stressed", or only level 1, or what? The perceptions must also be compared with the acoustic features, to be discussed next.

### 4.4 Acoustic Analysis

The spoken scripts were processed through the linear predictor, formant tracker, prosodic processors and boundary detector, described in section 3, to obtain data on the $F_0$ contours, intensity contours, formants, and other acoustic features that may correlate with stress, reduction, and syntactic boundaries. A variety of $F_0$-contour parameters and intensity-contour parameters are to be analyzed as potential correlates of stress. In this study, only $F_0$ contours are used in detecting constituent boundaries, in accord with the constituent boundary detector design described in section 3.4.

#### 4.4.1 $F_0$ Correlates of Stress

The $F_0$ parameters to be considered as possible acoustic cues to stress include: the peak $F_0$ in the vowel*; the mean $F_0$ in the vowel; the $F_0$ rise or fall in the initial portion of the vowel; the $F_0$ contour shape throughout the vowel (Medress, Skinner, and Anderson, 1971); and $F_0$ values in voiced consonants. Also to be considered are the coefficients of polynomial representations of $F_0$ contours (Levitt and Rabiner, 1971).

All such $F_0$ parameters are not easy to automatically extract from $F_0$ contours. $F_0$ peaks may occur in transitions between non-vowel sonorants and vowels, rather than within the vowel, and in such cases may or may not closely correspond with stress values. Also, $F_0$ peak values will depend upon the identity (specifi-

*Where "vowel" is used here and throughout this section, the syllable nucleus will also be considered (Lehiste, 1972; Stevens and Klatt, 1968; Stevens, 1969).

cally, the high/low feature) of the vowel, with low vowels showing lower $F_o$ peaks for the same laryngeal tensions and subglottal pressure. Average values are similarly affected, but they are even more difficult to automatically extract, since the durations over which the average must be computed must be determined, and an averaging computation is required. Finding the $F_o$ rise or fall in the initial portion of the vowel demands finding the vowel onset, which is particularly difficult following sonorants, but much easier following unvoiced consonants. $F_o$ contours throughout the vowel also require establishing the endpoints of the vowel. $F_o$ values in consonants require establishing the locations of consonants.

Since these contour parameters are highly dependent upon locating vowels and consonants and their boundaries, they may not lend themselves to easy automatic extraction, even if when manually extracted they correlate well with stress patterns. However, if some other features, such as intensity contours or formant structure, can locate vowels or consonants and their approximate boundaries, then these $F_o$ contour parameters may be machanically extracted.

### 4.4.2 Intensity Correlates of Stress

Intensity parameters to consider include: the maximum intensity in the vowel; the average intensity in the vowel; the intensity rise in the initial portion of the vowel; the integral of the energy within the vowel; the overall intensity contour shape in the vowel; energy within prevocalic and postvocalic consonants; and the presence of aspiration after stop releases. Full broadband energy will include aspirations, high-frequency frication noise, and stop bursts. Low-pass filtered energy may be used, however, to detect phonation or voicing energy, which would be high in vowels, presumably smaller in nonvocalic sonorants, low in voiced obstruents, and near zero in unvoiced consonants. This may provide come cues to the presence and locations of various categories of consonants and vowels.

Maximum intensity and average intensity in a vowel may correlate well with stress in neutral intonation positions, but the energy integral within a vowel, according to earlier studies (Lehto, 1971; Medress, Skinner, and Anderson, 1971), would be even better. However the energy integral requires determining the boundaries of the vowel, which is not always readily accomplished mechanically.

High-frequency energy within obstruents may be higher when they are in stressed syllables, so that the difference between total energy and phonation energy might be considered as a potential cue to stress.

### 4.4.3 Phonetic Durations as Stress Correlates

Duration and timing parameters that will be considered as potential stress cues include: duration of the portion of a vowel which has rising intensity (from previous consonant to energy peak in the vowel nucleus; cf. Lehto, 1969); total vowel duration; durations of prevocalic and postvocalic consonants; time intervals between vowel centers (peaks); and total syllabic durations.

To mark phonetic boundaries (and thus establish phonetic durations), significant changes in some acoustic features would presumably be sought. If the $F_o$ contours and intensity contours do not provide sudden changes appropriate for such boundary marking, then other acoustic features would have to be considered. Spectral structure (e.g. presence and positions of formant structures) might provide some cues. Ultimately, however, we recognize that speech is not a sequence of discrete acoustic units corresponding to individual phonemes, and such phonetic durations can, at best, be approximate indications of the quantity of the wave most closely associated with each phone. While subjectively determined durations (based on personal judgments as to where significant acoustic changes occur) may be found to correlate well with stress levels, the ease with which they may be mechanically determined will play a major role in determining their utility for speech recognition systems.

### 4.4.4 Vowel Quality and Reduction

Vowel quality is known to tend more toward schwa-like sounds for many weakly stressed and reduced syllables. Reduction often causes diphthongs to degenerate to single pure vowels (Fry, 1958), and causes general centralization of a vowel. Since unstressed and reduced vowels tend to be quite short, the vowel 'target' positions of articulation are not attained and thus formant target values are not reached (Lindblom, 1963). These effects may provide spectral features that can be correlated with stress levels.

4.4.5 Initial Examples

An earlier study of acoustic correlates of stress in isolated words (Medress, Skinner, and Anderson, 1971) showed that vowel duration was longer in (98% of all) stressed syllables than in unstressed syllables, when the stressed syllable was utterance-final. However, when the stressed syllable was in earlier portions of the utterance, durations were not as reliably correlated with stress (with the stressed vowel duration being longest in 67% of all cases where it is in a medial syllable, and only 51% of the time in stressed initial syllables). On the other hand, peak fundamental frequency was most reliably associated with stress in the earliest syllables (forming the intonation head for the utterance, in common intonation-contour terms; cf. Lehto, 1969). Peak $F_0$ was highest in the stressed syllable 93% of the time for initial syllables, 60% of the time for medial syllables, and 40% of the time for utterance-final syllables. Similarly, the average energy in the vowel was highest in the stressed syllables in 89%, 63%, and 60% of the initial, medial, and final syllables, respectively.

Thus these studies with isolated words showed effects that the position in the utterance had on the reliability of acoustic correlates of stress. Such effects may also be expected in connected speech. For example, in section 4.3, paragraph numbered (10), illustrations were given of such effects of position in the clause "they act like a prism" from the Rainbow Script (see also Figure 3).

4.5 Interpreting the Data

The prosodic data (as in Figure 3), and the digital spectrograms and formant tracks, all obtained for all six talkers reading the Rainbow Script, will provide extensive acoustic data to relate to the perceptual and linguistic data about stress patterns. The $F_0$ data processed through the constituent boundary detector will also yield boundary data to be compared with expected constituent boundaries.

The percentage of all expected syntactic boundaries that are correctly detected will be determined, as will the number of "false alarms" (where boundaries were not expected).

A syllable-by-syllable comparison of the acoustic correlates with perceived stress levels (probably based on the majority or unanimous judgments of all the listeners) will be conducted. The acoustic features and perceptual data will also be compared with the linguistically-predicted stress patterns.

When listeners agree as to the stress level of syllables spoken by a talker, these results may well be taken as the standard to compare with linguistic predictions and acoustic data. Where acoustic correlates or perceptions differ radically from linguistic predictions, the fault may be more in the gap between speech performance and linguistic models of competence, than in difficulties of perceptual or acoustic analysis.

Ultimately, the results of correlating acoustic, perceptual, and linguistic data may be difficult to precisely interpret because of several interfering factors, such as syntactic phrase structure, positions in sentence intonation contours, phonetic sequence effects, etc. Studies with the Rainbow Script will presumably indicate instances where such factors can or cannot be readily isolated. To isolate such factors more completely, sentences and connected texts will be specifically designed, as part of the further studies outlined in section 5.

# 5. FURTHER STUDIES

## 5.1 Reviewing Speech Texts for the ARPA Data Base

The Univac DSD Speech Communications Group is currently involved in a program to select good speech texts for use by the systems contractors of the Speech Understanding Research Program. In cooperation with Professor Michael O'Malley of the University of Michigan and Dr. June Shoup of Speech Communications Research Laboratory (SCRL), a two-phase effort is being undertaken. In one phase, sentences of general interest to the five systems contractors will be selected from a larger set which the system builders select as representative of the type data they hope to handle. In another phase of the data-base design, texts will be very carefully designed to isolate problems that are expected to be encountered in extendable speech recognition systems.

To accomplish phase one of the program, each system builder will select about fifty sentences of the type they expect their system to handle, giving consideration to problems they anticipate encountering in their system. These will be recorded and provided, along with an orthographic transcription and a list of criteria used to select the texts, to SCRL, the Univeristy of Michigan, and Univac, for their review.

The Univac Speech Communications Group will give particular attention to evaluating characteristics of the data that relate to prosodic patterns (such as sentence intonation contour, position in discourse, and stress patterns) and higher-level linguistic considerations (number of words in the vocabulary, variety of words in each parts-of-speech category, representative variety of syntactic structures, sentence types, semantic relationships, etc.). Other factors being considered (primarily by the other two groups) include phonemic variety and balance, allophonic variations in the text, morphophonemic phenomena (such as coarticulation effects, word or morpheme boundaries, etc.), dialect differences, and style (read speech, spontaneous speech, etc.).

Following the separate evaluations of some of these characteristics by each

of the three groups, a common workshop will be held where the three groups, in cooperation with James Forgie of Lincoln Laboratory, will select, from the 250 sentences, ten good sentences to be recorded later by two talkers from each systems contractor, plus 50 to 100 other sentences, either selected from the remainder of the acceptable portion of the original 250 sentences, or specifically designed to fill any gaps in the data. These selected utterances will form an initial part of the Lincoln Speech Data Facility.

Later, the three groups and systems contractors will meet to devise specific ways in which prosodic features and phonetic patterns in the selected data may be used in each of the five systems. Univac plans to process some of the data through the constituent boundary detection program, and to estimate (or actually obtain) perceptual and acoustic data about stress patterns in representative utterances. Such indications of actual or expected prosodic patterns will be available for system builders and others to use in devising prosodic aids to speech understanding systems.

## 5.2 Designing Sentences for Isolating Prosodic Effects

Phase two of the data base design program is concerned with designing a set of sentences or texts which isolate certain factors which may affect the success of speech understanding systems.

One way to determine what is causing any particular pattern in speech data (such as the presence or absence of an $F_0$ valley at a constituent boundary, or the occurence of a long or short vowel duration in a stressed syllable in utterance-final position) is to compare utterances which are similar except for only one or a few differences. This is how phonemes of English are sometimes determined, using minimal pairs (such as pit/pet for distinguishing /I/ and /ɛ/). Likewise, acoustic correlates of stress have been studied based on contrasting words like permit versus permit, in which only the stress patterns differ in the two words. These techniques may be applied to determining effects of different syntactic stuctures, sentence types, intonation contours, phonemic structure, etc.

The Univac Speech Communications Group is designing a set of utterances which can isolate the specific effects (particularly on prosodic patterns) of some of the following factors: spontaneous speech versus spoken texts versus speech actually used in man-machine interaction; positions of sentences or other speech portions within paragraphs or discourse; the type of sentence (declarative, yes-no question, WH-question, or command); positions of phrases within an overall sentence intonation contour (intonation body or tonalization positions); simplex versus complex sentences; special sentence transformations (e.g., adverb preposing or passives); special phrase category effects (e.g., unstressed nature of pronouns); vocabularies and subvocabularies; phonetic variety and balance (dc both voiced and unvoiced consonants follow a vowel, etc.); and effects of error in pronunciation and analysis (which analysis errors may give the most lexical errors in bottom-up analyzers, etc.).

With so many dimensions to be independently controlled and tested, the difficult problem is how to keep the set of designed utterances to a reasonable size. Among techniques being incorporated to restrict the data set is the obvious procedure of incorporating within a single sentence several contrasts that are still distinguishable and isolated (by their distance apart in the sentence, for example). Another procedure is to first study those contrasts which are most likely to affect performance of speech understanding systems, leaving subtleties for later study.

## 5.3 Guidelines to Use of Prosodies in Speech Understanding Systems

As part of a continuing effort in coordinating the Univac studies with activities of other ARPA contractors, Wayne Lea is preparing a tutorial for presentation at a forthcoming meeting of ARPA contractors. This tutorial deals with aspects of prosodies, syntactic structure, and semantics which are of interest to systems builders. It is being coordinated with other presentations to be given by Dennis Klatt of Massachusetts Linstitute of Technology, Mike O'Malley of the University of Michigan, and June Shoup of Speech Communications Research Laboratory. These tutorials will collectively summarize many aspects of acoustic processing, acoustic phonetics, phonemics, morphophonemics, prosodies, syntax, semantics, and pragmatics.

Plans are also being made for using prosodic features to aid distinctive-features estimation routines, syntactic parsers, and semantic processors, using the ARPANET for access to programs at facilities of other ARPA contractors.

In general, these efforts with prosodic aids to speech understanding systems are part of a general strategy to use the most reliable information in the acoustic data in conjunction with early use of linguistic regularities. Prosodic features are expected to play a crucial role in such a strategy for sentence recognition. Their effectiveness will depend upon how well they are incorporated into total systems for speech understanding.

## 6. CONCLUSIONS

Research on prosodic aids to speech recognition is still in progress. Conclusions at this point thus necessarily are confined to a synopsis of the general strategy and motivation of the work, and a few preliminary judgments from the limited stress perception data and boundary detection results.

Linguistic and perceptual arguments clearly suggest the value of early use of syntactic hypotheses in recognition routines. Prosodic features can provide cues to the presence of syntactic constituent boundaries and to the stress pattern of the spoken utterance. In particular, fall-rise patterns in voice fundamental frequency contours mark major constituent boundaries. Sentence boundaries are usually marked by pauses followed by large increases in fundamental frequency in the beginnings of new sentences. While prosodic features of fundamental frequency, intensity, and phonetic durations are known to be associated with English stress levels in isolated utterances, the best acoustic correlates to use in automatically determining stress levels in connected speech must still be found. Techniques for using prosodic features in aiding distinctive features estimation, syntactic parsing, and semantic representation are yet to be implemented and tested.

At Univac DSD, the basic strategy for acoustic speech analysis is to locate the reliable, clearly-encoded prosodic and distinctive features in the acoustic data, and incorporate them immediately with linguistic regularities to provide the data for generating syntactic hypotheses, lexical identifications, and semantic judgments. One general procedure being considered is to locate boundaries between major grammatical constituents, find the stressed syllable(s) in each constituent, and do a partial distinctive features analysis within the stressed syllables, where distinctive features are expected to be most clearly and consistently manifested.

The program for detecting constituent boundaries from fundamental frequency contours, as implemented at Univac, appears to give satisfactory results, although it must still be tested with further texts, including questions and commands, and it may profit from refinements to eliminate false alarms at some consonant-vowel

boundaries and to incorporate energy cues within the algorithm.

Basis analysis tools needed for the partial distinctive features estimation have been implemented. These include linear predictor analysis and formant tracking. Preliminary attempts at restricted distinctive features estimation have been incorporated into previous large-vocabulary word-recognition or phrase-recognition schemes (Medress, 1972), but much more is to be done to devise adequate distinctive features estimation schemes suitable for use in the stressed syllables of connected speech.

To find the stressed syllables wherein the major distinctive features estimation effort will be concentrated, more must be learned about stressed syllables in connected speech. The experiments with the Rainbow Script are designed to interrelate linguistic predictions, perceptual judgments, and acoustic data about stressed, unstressed, and reduced syllables in connected speech. A syntactic analysis and subsequent application of published stress rules will yield testable predictions about normative stress patterns for the script. These will be compared with perceptual judgments and with acoustic features, including several parameters of fundamental frequency contours, energy contour parameters, durations, and vowel quality. However, considering the frequently discussed gap between ideal linguistic competence and actual listener or talker performance, the theoretical predictions are not to be taken as the standard for determining acoustic correlates of stress. The performance of listeners who judge stress level based on what they hear in the speech signal may be argued to be a better standard for assessing acoustic correlates.

As of October, 1972, the acoustic data (contours of fundamental frequency and energy, digital spectrograms, and formant tracks) had been obtained for six talkers reading the Rainbow Script. The syntactic analysis had not been done, but some of the perceptual data had been gathered.

The partial results from some listeners' judgments as to whether syllables are stressed, unstressed, or reduced suggest several preliminary conclusions. When listeners heard clausal portions of the text repeated at will (by tape rewind

and replay) or digitized, stored, and repetitively retrieved and D/A converted, they were able to distinguish individual stressed, unstressed, and reduced syllables. The close correspondence between the average performance of five listeners and the performance of one listener (WAL) showed that listener to be representative of the set of listeners.

Listener WAL performed the test for all six talkers under three conditions: (1) when tape rewind was used and only "stressed" and "reduced" syllables were marked (making those unmarked be "unstressed" by default; cf. Hughes, Li, and Snow, 1972); (2) when the digitizing and computer replay method was used, and each syllable was marked as either "highly stressed", "lesser stressed" unstressed, or reduced; and (3) when each syllable must be marked as stressed, unstressed, or reduced, and the tape rewind method was used. The three conditions gave similar results for syllable-by-syllable judgments, except that when each syllable was not necessarily marked, many syllables were apparently "unstressed" by default (that is, they perhaps should have been marked reduced or stressed, but they were "missed" in the process of marking only stressed and reduced syllables). Most "highly stress" and "lesser stressed" syllables from the computer-replay run were judged as "stressed" in the three-level tape-replay run.

Results did differ somewhat from talker to talker, and are affected by the individual listeners' judgments, as might be expected. Yet, there was considerable agreement about the stress levels of many syllables, regardless of talker or listener (cf. Hughes, Li, and Snow, 1972). For example, when listener WAL marked stressed, unstressed, and reduced syllables for each talker, using the tape rewind method, the total results of comparing the other five talkers to ASH (as in Figure 9f, page 49) showed that only 16% of all judgments for ASH differed from those for any of the other talkers. Over two thirds of these "confusions" were between unstressed and reduced syllables.

Of the 127 syllables in the Rainbow Script, about 50 to 60% were judged by listener WAL to be stressed, (depending upon the talker), about 35 to 40% were judged as reduced, and 10 to 25% were judged as unstressed. The strategy of doing a partial distinctive features analysis in only stressed syllables thus

would lighten the load on acoustic analysis considerably, while still allowing acoustic analysis on enough syllables that considerable distinctive features data can be available for guiding lexical hypothesis making and aiding higher-level structural analyses.

Further perception tests are to be made, with other listeners marking all syllables, and with repetition tests by one listener under identical conditions.

The Rainbow Script is recognized as not being ideal for isolating and studying the individual effects of intonation contour, position in the sentence, phrase structure, semantic structure, and phonetic content on stress and boundary results. The design of texts to isolate such factors is being undertaken. In addition, Univac will be evaluating whether speech data recorded by ARPA systems contractors will demonstrate systems effectiveness under varieties of such prosodic, syntactic, phonemic, and semantic conditions. Efforts will be undertaken to integrate prosodic information into other programs on total speech understanding systems.

## 7. REFERENCES

ALLEN, J. (1968), A Study of the Specification of Prosodic Features of Speech from the Grammatical Analysis of Printed Text, Ph.D. Thesis, Dept. of E.E., M.I.T.

ARMSTRONG, L. E. and WARD, I. C. (1926), Handbook of English Intonation. Cambridge: Heffer (2nd Edit.).

BARIK, H. C. (1969), On Defining Juncture Pauses: A note on Boomer's "Hesitation and Grammatical Encoding", Language and Speech, vol. 11, pp. 156-159.

BARNWELL, T. P. (1970), Initial Studies on the Acoustic Correlation of Prosodic Features for a Reading Machine. QPR No. 93, Research Laboratory of Electronics, M.I.T., pp. 262-271, Also TR 749, M.I.T., R.L.E.

BERMAN, A., and SZAMOSI, M. (1972), Observations on Sentential Stress, Language, vol. 48, 304-325.

BEVER, T. G., LACKNER, J. F., and KIRK, R. (1969), The Underlying Structures of Sentences Are the Primary Units of Immediate Speech Processing, Perception and Psychophysics, vol. 5, pp. 225-34.

BIERWISCH, M. (1965), Regeln fur die Intonation deutscher Satze, Studia Grammatica, vol. 7, pp. 99-201.

BLACKMAN, R. B., and TUKEY, J. W., (1958) The Measurement of Power Spectra, Dover Publication Inc., New York.

BLESSER, B. (1969), Perception of Spectrally Rotated Speech, Ph.D. Thesis, E.E. Dept., M.I.T.

BOLINGER, D. (1958), A Theory of Pitch Accent in English. Word, vol. 14, p. 109.

BOOMER, D. S. (1965), Hesitation and Grammatical Encoding, Language and Speech, vol. 8, pp. 148-158.

BRESNAN, J. (1971), Sentence Stress and Syntactic Transformations, Language, vol. 47, pp. 157-81.

BRESNAN, J. (1972), Stress and Syntax: A Reply, Language, vol. 48, pp. 326-42.

BURT, M. K. (1971), From Deep to Surface Structure: An Introduction to Transformational Syntax. New York: Harper and Row.

CANTRELL, W. R. (1969), Pitch, Stress, and Grammatical Relations. Papers from the Fifth Regional Meeting of the Chicago Linguistic Society, Chicago: Univ. of Chicago Press, pp. 12-24.

CHOMSKY, N. (1964), Current Issues in Linguistics. In The Structure of Language (J. Fodor and J. Katz, Eds.), pp. 50-118.

CHOMSKY, N. (1965), Aspects of the Theory of Syntax. Cambridge, Mass: M.I.T. Press.

CHOMSKY, N. and HALLE, M. (1968), The Sound Pattern of English. New York: Harper and Row.

CHOMSKY, N. and MILLER, G. A. (1963), Introduction to the Formal Analysis of Natural Languages. In Handbook of Mathematical Psychology, pp. 269-321. Ed. R. D. Luce, R. R. Bush and E. Galanter, New York: John Wiley and Sons, Inc.

CRYSTAL, D. (1969), Prosodic Systems and Intonation in English. Cambridge: Univ. Press.

DELATTRE, P. (1965), Comparing the Phonetic Features of English, French, German, and Spanish. Heidelberg: Julius Gross Verlag.

FAIRBANKS, G. (1940), Voice and Articulation Drillbook. New York: Harper and Row.

FLANAGAN, JAMES L. (1965) Speech Analysis, Synthesis and Perception. New York: Academic Press.

FODOR, J. A. and GARRETT, M. (1966) Some Reflections on Competence and Performance. In Psycholinguistics Papers (J. Lyons and R. J. Wales, Eds.), pp. 135-54. Edinburgh: Edin. U.P.

FRY, D. B. (1955), Duration and Intensity as Physical Correlates of Linguistic Stress. J. Acoust. Soc. Amer., vol. 35, pp. 765-769.

FRY, D. B. (1958), Experiments in the Perception of Stress. Language and Speech, vol. 1, pp. 126-152.

GLEASON, H. A. (1961), An Introduction to Descriptive Linguistics. New York: Holt, Rinehart, and Winston.

GLEITMAN, L. R. and GLEITMAN, H. (1970) Phrase and Paraphrase. New York: W. W. Norton and Co.

GOLDMAN-EISLER, F. (1958), The Predictability of Words in Context and the Length of Pause in Speech, Language and Speech, vol. 1, pp. 226-231.

GOLDMAN-EISLER, F. (1961), A Comparative Study of Two Hesitation Phenomena, Language and Speech, vol. 4, pp. 18-26.

GRIMES, J. E. (1969) Phonological Analysis: Part One. Santa Ana, Calif.: Summer Inst. Ling.

HAGGARD, M. P. (1967) Models and Data in Speech Perception. In Models for the Perception of Speech and Visual Form (W. Wathen-Dunn, Ed.), pp. 331-8, Cambridge, Mass.: M.I.T. Press.

HALLE, M. and KEYSER, S. J. (1971), English Stress. New York: Harper and Row.

HOGG, R. V. and CRAIG, A. T. (1965), Introduction to Mathematical Statistics, the Macmillan Company, New York. (p. 62)

HOUSE, A. S. (1960) On Vowel Duration in English, J. Acoust. Soc. Amer., vol. 33, pp. 1174-1178.

HOUSE, A. S. and FAIRBANKS, G. (1953) The Influence of Consonant Environment upon the Secondary Acoustical Characteristics of Vowels, J. Acoust. Soc. Amer., vol. 25, pp. 105-113.

HUGHES, G. W., LI, K.-P., and SNOW, T. B. (1972), An Approach to Research on Word Spotting in Continuous Speech, Proc. 1972 Conf. on Speech Communication and Processing, Newton, Mass., pp. 109-112.

HULTZEN, L. S. (1957) Communication in Intonation: General American. Study of Sounds, Phonetics Soc., Japan. Tokyo, pp. 317-333.

HULTZEN, L. S. (1959) Information Points in Intonation. Phonetica, vol. 4, pp. 107-120.

HUTTAR, G. L. (1968), Two Functions of the Prosodics in Speech. Phonetica, vol. 18, pp. 23 -241.

JONES, D. (1909) Intonation Curves. B. G. Teubner, Leipzig and Berlin.

JONES, D. (1932) Outline of English Phonetics. Cambridge: Heffer (8th Ed.)

KEYSER, S. J. (1969) The Linguistic Basis of English Prosody. In Modern Studies in English (Reibel and Schane, Eds.), Englewood Cliffs, New Jersey: Prentice-Hall, pp. 379-394.

LAKOFF, G. (1972), The Global Nature of the Nuclear Stress Rule, Language, vol. 48, 285-303.

LEA, W. A. (1968), Establishing the Value of Voice Communication with Computers. IEEE Trans. Audio Electroacoustics, AU-16(2), pp. 184-197.

LEA, W. A. (1970), Towards Versatile Speech Communication with Computers, Intern. J. Man-Machine Studies, vol. 2, pp. 107-155.

LEA, W. A. (1971), A Formalization of Measurement Scale Forms. Journal of Math. Sociology, Vol. 1, 81-104.

LEA, W. A. (1972a), Intonational Cues to the Constituent Structure and Phonemics of Spoken English, Ph.D. Thesis, School of E.E., Purdue University.

LEA, W. A. (1972b), An Approach to Syntactic Recognition without Phonemics. Proc. 1972 Intern. Conf. on Speech Commun. and Processing. Newton, Mass.: pp. 198-201. A revised version of this paper has been accepted for publication in the IEEE Transactions on Audio and Electroacoustics.

LEA, W. A. (IN PRESS) Computer Recognition of Speech: Current Trends in Linguistics, The Hague, Netherlands: Mouton and Co., 1561-1620.

LEA, W. A. and LI, K.-P. (IN PREPARATION) Experiments on Stress and Constituent Structure of Spoken English.

LEHISTE, I. (1970), Suprasegmentals. Cambridge: M.I.T. Press.

LEHISTE, I. (1972), The Syllable Nucleus as a Unit of Timing, J. Acoust. Soc. Amer., vol. 52, No. 1, p. 182.

LEHTO, L. (1969), English Stress and its Modification by Intonation. (Suomalaisen Tiedcakatemian Toimituksia, sarja B, Mide 196). Helsinki: Academia Scientiarum Fennica.

LEOPOLD, W. F. (1953), Patterning in Children's Language, Language Learning, vol. 5, 1-14.

LEVITT, H. and RABINER, L. (1971), Analysis of Fundamental Frequency Contours in Speech. J. Acoustical Soc. Amer., vol. 49, pp. 569-82.

LEWIS, M. (1936), Infant Speech, A Study of the Beginnings of Language. New York: Harcourt Brace.

LIEBERMAN, P. (1960), Some Acoustic Correlates of Word Stress in American English. J. Acoust. Soc. Amer., vol. 32, pp. 451-454.

LIEBERMAN, P. (1965) On the Acoustic Basis of the Perception of Intonation by Linguists. Word, vol. 21, pp. 40-54.

LIEBERMAN, P. (1967a), Intonation, Perception, and Language. Cambridge: M.I.T. Press.

LIEBERMAN, P. (1967b), Intonation and Syntactic Processing of Speech. Models of the Perception of Speech and Visual Form (W. Wathen-Dunn, Ed.), Cambridge, Mass.: M.I.T. Press.

LINDBLOM, B. (1963) "Spectrographic Study of Vowel Reduction" J. Acoust. Soc. Amer., vol. 35, No. 11.

LUMMIS, R. C. (1971), Real-Time Technique for Speaker Verification by Computer, J. Acoust. Soc. Amer., vol. 50, p. 106.

MAKHOUL, J. (1972), Aspects of Linear Prediction in the Spectral Analysis of Speech, Proc. 1972 Conf. on Speech Communication and Processing, Newton, Mass. pp. 77-81.

MAKHOUL, J. and WOLF, J. (1972), Linear Prediction and the Spectral Analysis of Speech, Technical Report, ARPA Contract No. DAHC-71-C-0088, Bolt Bernanek and Newman Report No. 2304.

MALMBERG, B. (1963) Structural Linguistics and Human Communication. New York: Academic Press.

MATTINGLY, I. (1966), Synthesis by Rule of Prosodic Features. Lang. and Speech, vol. 9, pp. 1-13.

MEDRESS, M. (1969), Computer Recognition of Single-Syllable English Words. Ph.D. Thesis, Dept. of E. E., M.I.T.

MEDRESS, M. (1972) A Procedure for the Machine Recognition of Speech, Proc. 1972 Conf. on Speech Communication and Processing, Newton, Mass., pp. 113-116, April.

MEDRESS, M. F. and SKINNER, T. E. (1972), Acoustic Correlates of Linguistic Stress. SUR Note #36, ARPA NIC 10524.

MEDRESS, M. F., and SKINNER, T. E., and ANDERSON, D. E. (1971), Acoustic Correlates of Word Stress, Presented to 82nd Meeting, Acoustical Society of America, Denver, Colorado, October 20, (Paper K3).

MILLER, G. A. (1962), Decision Units in the Perception of Speech, IRE Trans. on Info. Th., vol. IT-8, pp. 81-3.

MOL, H. and UHLENBECK, E. M. (1956), The Linguistic Relevance of Intensity in Stress, Lingua, vol. v, 205-13.

MORTON, J. and JASSEM, W. (1965), Acoustic Correlates of Stress, Lang. and Speech, vol. 8, 159-81.

NOLL, A. M. (1967), Cepstral Pitch Determination, J. Acoust. Soc. Amer., vol. 41, pp. 293-309.

O'MALLEY, M. H. and PETERSON (1966), An Experimental Method for Prosodic Analysis, Phonetica, vol. 15, Number 1, pp. 1-13.

O'MALLEY, M. H. (1972), Personal communication.

PETERSON, G. (1963), Automatic Speech Recognition Procedures. In Automatic Speech Recognition, vol. II, pp. Ha83-Ha102. Ann Arbor: Univ. of Michigan Press.

PIKE, K. L. (1945), The Intonation of American English. Ann Arbor: University of Michigan.

ROSS, J. R. (1969), A Reanalysis of English Word Stress - Part I. Unpublished manuscript, M.I.T.

SCHOLES, R. J. (1971) Acoustic Cues for Constituent Structure. The Hague: Mouton and Co.

SHAFER, R. W. and RABINER, L. R. (1970), System for Automatic Formant Analysis of Voiced Speech, J. Acoust. Soc. Amer., vol. 47, pp. 634-648.

SONDHI, M. M. (1968), New Methods of Pitch Extraction, IEEE Trans. on Audio and Electroacoustics, vol. AU-16, pp. 262-266.

STEVENS, K. N. (1969), Study of Acoustic Properties of Speech Sounds II, and Some Remarks on the Use of Acoustic Data in Schemes for Machine Recognition of Speech. Scientific Report No. 12, Contract No. F19628-68-C-0125, report AFCRL-69-0339, prepared by Bolt, Beranek, and Newman, Cambridge, Mass.

STEVENS, K. N. (1971), Perception of Phonetic Segments: Evidence from Phonology, Acoustics, and Psychoacoustics. Copies available from the Research Laboratory of Electronics, M.I.T.

STEVENS, K. N., and KLATT, M. (1969), Study of Acoustic Properties of Speech Sounds. Scientific Report No. 8, Contract No. F19628-68-C-0125, report AFCRL-68-0446, prepared by Bolt, Beranek, and Newman, Cambridge, Mass.

STEVENS, S. S. (1951), Mathematics, Measurement, and Psychophysics. In Handbook of Experimental Psychology (S. S. Stevens, Ed.). New York: John Wiley and Sons, 1951, 1-49.

STOCKWELL, R. P. (1960) The Place of Intonation in a Generative Grammar of English, Language, vol. 36, p. 360.

TRAGER, G. L. and SMITH, H. L., JR. (1951), An Outline of English Structure, Studies in Linguistics: Occasional Papers 3, Norman, Oklahoma: Battenburg Press.

TRUBETZKOY, N. S. (1969), Principles of Phonology. Berkeley: U. Calif. Press, (From the French of 1939).

VANDERSLICE, R. (1968) The Prosodic Component: Lacuna in Transformational Theory. Report P-3674, the RAND Corp.

VANDERSLICE, R. and LADEFOGED, P. (1971), Binary Suprasegmental Features. Working Papers in Phonetics No. 17, Phonetics Lab., Univ. of Calif. at Los Angeles, pp. 6-23.

WELLS, R. S. (1947), Immediate Constituents. Language, vol. 23, pp. 81-117.

WILLEMS, Y. (1972), The Use of Prosodics in the Automatic Recognition of Spoken English Words. Ph.D. Thesis, Dept. of E.E., M.I.T.